The Dissertation Committee for Shu Xu
certifies that this is the approved version of the following dissertation:

Data cleaning and knowledge discovery in process data

Committee:

Thomas F. Edgar, Supervisor

Willy Wojsznis

Dragan Djurdjanovic

Gary T. Rochelle

Michael Baldea

Michael J. Daniels

Data cleaning and knowledge discovery in process data

by

Shu Xu, B.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2015

Dedicated to my dear parents Jinluo Xu and Chunru Wang.

Acknowledgments

Foremost, I want to thank my advisor, Professor *Thomas F. Edgar*, for his trust, patience, availability, and broad knowledge guided me throughout my PhD life. I would also like to thank those outstanding professors and engineers for serving on my thesis committee: *Dragon Djurdjanovic, Gary T. Rochelle, Michael Baldea, Willy Wojsznis*, and *Michael Daniels*.

I was also fortunate enough to study together with many exceptional graduate students including: *Bo Lu, Siyun Wang, Matt Walters, Victor C. Duribe, Jungup Park, Ankur Kumar, Ray Wang, Abigail Ondeck* and *Ricardo Dunia*. This is an excellent group of people to be associated with from both academia and personal perspective.

Last but not the least, I would like to thank engineers from Emerson Process Management: *Terrence Blevins, Mark Nixon*, for supporting my research and sharing valuable industrial perspective and experience. I enjoyed the collaboration from the bottom of my heart.

Shu Xu

The University of Texas at Austin
October 2015

Data cleaning and knowledge discovery in process data

Publication No. _____

Shu Xu, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Thomas F. Edgar

This dissertation presents several methods for overcoming the Big Data challenges, with an emphasis on data cleaning and knowledge discovery in process data. Data cleaning and knowledge discovery is chosen as a main research area here due to its importance from both theoretical and practical points of view.

Theoretical background and recent developments of data cleaning methods are reviewed from four aspects: missing data imputation, outlier detection, noise removal and time delay estimation. Moreover, the impact of contaminated data on model performance and corresponding improvement obtained by data cleaning methods are analyzed through both simulated and industrial case studies. The results provide a starting point for further advanced methodology development.

It is hard to find a universally applicable method for data cleaning since every data set may have its own distinctive features. Thus, we have to customize available methods so that the quality of the data set is guaranteed. An

integrated data cleaning scheme is proposed, which incorporates model building and performance evaluation, to provide guidance in tuning the parameters of data cleaning methods and prevent "over-cleaning". A case study based on industrial data has been used to verify the feasibility and effectiveness of the proposed new method, during which a partial least squares (PLS) model was built and three univariate data cleaning procedures is tested.

A time series Kalman filter (TSKF) is proposed that successfully handles outlier detection in dynamic systems, where normal process changes often mask the existence of outliers. The TSKF method combines a time series model fitting procedure with a modified Kalman filter to deal with additive outlier (AO) and innovational outlier (IO) detection problems in dynamic process data set. A comparative analysis of TSKF and available methods is performed on simulated and real chemical plant data.

Root cause diagnosis of plant-wide oscillations, as a concrete example of data cleaning and knowledge discovery in the process data, is provided. Plant-wide oscillations can negatively influence the overall control performance of the process and the detection results are often affected by noise at different frequency ranges. To address such a problem, an information transfer method combining spectral envelope algorithm with spectral transfer entropy is proposed to detect and diagnose such oscillations within a specific frequency range, mitigating the effects from measurement noise. The feasibility and effectiveness of the proposed method are verified and compared with available methods through both simulated and industrial case studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Big data challenge

Recently, the emerging data related problems represented by the term "Big Data"have challenged both current databases software tools and data scientists [199]. Big Data are characterized with volume, variety and velocity, or simply V^3 [350]: the volume of data sets changes between terabytes (10^{12} bytes) and zettabytes(10^{21} bytes); data sets contain various structures: process measurements, text, audio and images [306]; the streaming data obtained from continuous operation challenge the engineers' capabilities of performing on-line data analysis instead of off-line.

The smart manufacturing era [77] has witnessed IP-enabled intelligent devices such as wireless controllers and sensors being massively instrumented in the process industries [38] recording real-time process information at a high frequency. To handle such abundant process data, data-driven methods [290, 154] outperform physical-model based ones in performing tasks such as dimensionality reduction, variable selection [67, 207, 191], process monitoring, and fault detection[168, 195, 64]. For example, the partial least squares (PLS) algorithm provides process engineers with a powerful tool to quantitatively analyze the

near infrared (NIR) spectra during process monitoring [40].

However, data collected from plants usually suffer from lower quality problems, and they may contain missing values, outliers, noise and multi-level structures. For example, sampling rates and correlation levels may differ significantly for variables within the same unit and between different units [110]. Thus, we have to either clean the data before extracting knowledge from them[154], or we can make certain improvements on the methods' robustness when dealing with contaminated data [238].

1.2 Outline of this work

Since data cleaning serves as a significant step in the knowledge discovering process, it is necessary to explore fast and effective data cleaning methods to ensure data quality for further analysis [122].

In Chapter 2, data cleaning methods from not only the traditional applied statistics, but also the state-of-art machine learning discipline will be reviewed. The methods cover four different aspects: missing data imputation, outlier detection, noise removal and time delay estimation. In the end of Chapter 2, based on the literature review, we provide suggestions on evaluating data cleaning methods and pinpoint challenges and directions for researchers. The contributions of this dissertation are focused on the following four aspects: (1) how data cleaning impacts model performance; (2) how to combine data cleaning method with model performance evaluation; (3) how to improve the outlier detection performance when dealing with a dynamic data set; (4) how

to effectively dampen the noise effect, extract useful information from the data set and use it in process monitoring in frequency domain. To solve these challenges, the dissertation is divided into four chapters and a brief description of each chapter is shown as follows:

Chapter 3 studies the impact of outliers and noise on dynamic model identification and time delays on partial least squares (PLS) model performance. Data from an industrial process and a simulated case are used for the first and second scenarios. The results validate the argument that contaminated data sets negatively affect the model performance and it is necessary to incorporate data cleaning techniques in such a knowledge discovery process.

Based on the analysis provided on Chapter 3, an integrated data cleaning scheme is proposed in Chapter 4 which combines the data cleaning step with a later stage of model building and performance evaluation. The new scheme is compared with a robust version of partial least squares (RPLS) through simulated and industrial case studies and show satisfactory performances.

Besides concerns on model performance, intrinsic properties of the data sets should not be neglected when perform data cleaning, as mentioned in Chapter 2. For example, the process dynamics contained in the data sets might mask the existence of outliers. To deal with such a problem, a time series Kalman filter (TSKF) is devised in Chapter 5 which approximate dynamic variations with time-series models and detect the outliers based on the inconsistencies between model predictions and measurements. Despite the

computation cost concerns, the TSKF method can be applied both on-line and off-line for univariate and multivariate outlier detections, and obtained satisfactory results based on both simulated and plant data testing.

Chapter 6 provides a knowledge discovery example in the process industries– studying the information transfer between variables in the plant in the frequency domain. By implementing the spectral envelope method, the dominant frequency contained in the data set is unveiled and the noise effects are dampened. A spectral transfer entropy method is proposed which can be used in calculating the entropy transfer between variables at the dominant frequency. Such strategies are successfully applied in plant-wide oscillation detection and root cause diagnosis shown by the case studies.

Finally, conclusions of this work will be presented in Chapter 7 in which the contributions and future directions are indicated.

Chapter 2

Overview of data cleaning methods

As mentioned in Chapter 1, data cleaning is critical in the knowledge discovery process and it is necessary to explore fast and effective ways to ensure data quality. To get foundational knowledge about such an area, this chapter provides literature review on current data cleaning methods from the perspectives of missing data imputation, outlier detection, noise removal and time delay estimation, as shown in Sections 2.2 $\sim$ 2.5[1]. Moreover, a brief introduction to machine learning is provided to facilitate the understanding of those state-of-art data cleaning techniques. Based on the literature review, concluding remarks and suggestions are given in Section 2.6.

2.1 A breif introduction to machine learning

Machine learning originates from computer science – focusing on "teaching" machines to "learn" from existing data and guiding exploratory data analysis (EDA) [217, 51] and data mining [211, 35]. Data cleaning can benefit sig-

[1]Xu S, Lu B, Baldea M, Edgar TF, Wojsznis W, Blevins T, Nixon M. Data cleaning in the process industries. Reviews in Chemical Engineering. 2015; 31(5), 453-490.The project was supervised by Dr. Michael Baldea and Dr. Thomas F. Edgar. Willy Wojsznis, Terry Blevins and Mark Nixon gave technical support and conceptual advice

nificantly from machine learning algorithms because the later provide powerful tools to extract information from available observations that helps improve the data quality: imputing missing values, detecting outliers, etc.

Generally, machine learning algorithms can be categorized as follows:

- **Supervised learning**: predicting outputs from inputs based on a function inferred from a labeled training data set containing satisfactory inputs and outputs. Supervised learning includes classification and regression, and the corresponding algorithms include naive Bayes classifier [261], decision trees [251], k-nearest neighbor(kNN) [12], support vector machine (SVM) [70], regression (linear, logistic,...) and so on.

- **Unsupervised learning**: extracting patterns (or hidden structure) from the unlabeled training data without specifying an output value (or label). Clustering problems are generally considered as unsupervised learning, and available algorithms include density-based spatial clustering of applications with noise (DBSCAN) [97] , mean shift [106], k-means clustering [124] and so on.

2.2 Missing data imputation

2.2.1 Motivation

Missing values in process industries refer to entries in the data set that have no connection with the real state of the process and take values such as $\pm\infty$, 0, nan (not a number). Although it seems that the redundancy provided

by a large number of sensors makes missing data no longer a less important issue, imputing the missing values at the sensor (local) level may not guarantee satisfactory model performance, especially when the percentage of missing data is large and the correlations between variables are affected. Usually the missing value imputation is done by manual screening, but with abundant data newly generated, it is not feasible to do it by hand and we have to seek automated ways. The first step for process engineers to find an appropriate method is to investigate whether the pattern behind missing values is random or not and possible causes. Commonly faced missing patterns and related causes are shown in Fig. 2.1: $Y_1 - Y_5$ are process variables, a rectangle stands for measured data points, and the missing data are represented by white spaces between rectangles.

1. Only one variable (Y_5) contains missing values, perhaps due to a single sensor failure.

2. Observations of variables ($Y_2 - Y_5$) are missing for the same time stamps, probably indicating a fault condition occurred in a unit operation leading to non-responses.

3. Missing values irregularly show up, and this situation is caused by outlier removal and sensor breakdown or malfunction.

4. A single variable (Y_3) shows a regular missing pattern, which is generated by multi-rate sampling.

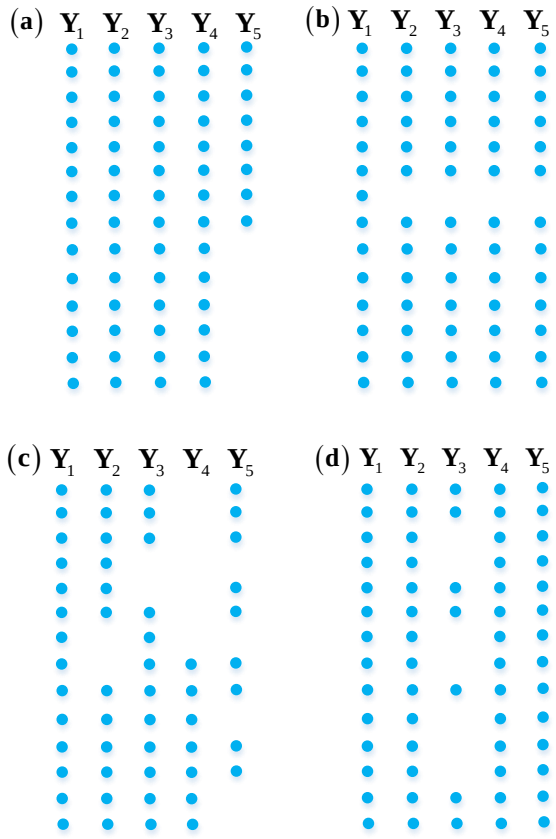


Figure 2.1: Common missing patterns in the process industries

A complete data set Y can be partitioned into subsets of observed data and missing data: $Y = (Y_{obs}, Y_{mis})$. Define R as a missingness indicator: if Y is observed, $R = 1$; otherwise, $R = 0$. There are generally three missing patterns [185]:

- **Missing completely at random(MCAR)**:the probability that an element of the data set is missing depends on neither the observed data nor the missing ones, in other words:

$$P(R|Y) = P(R) \tag{2.1}$$

  where $P(R|Y)$ is the conditional distribution of R given Y.

- **Missing at random (MAR)**: the probability that an element of the data set is missing depends on the observed data only, or:

$$P(R|Y) = P(R|Y_{obs}) \tag{2.2}$$

- **Missing not at random(MNAR)**: the probability that an element of the data set is missing depends on both the observed data and the missing data, or

$$P(R|Y) = P(R|Y) \tag{2.3}$$

Two methods can be used in diagnosis of missing patterns [185, 299]. The first method is to examine whether the pattern of group differences of Y_{mis} and Y_{obs} for a single variable exists for other variables of interest; if a

consistent pattern is found, then we can assume that the data are missing not at random. The second method is to calculate missing data correlation for variable pairs. After transforming the original data set into a dataset of binary numbers (1 for observed and 0 for missing), the correlation between the missing values on each pair of variables can be calculated. Statistical tests can be applied to evaluate the correlation and estimate the degree of randomness. Generally, if the randomness is significant for all variable pairs, we can assume that the data is missing completely at random (MCAR). However, in most cases, only some pairs of variables give high values of randomness, thus we can treat the data set as missing at random (MAR), which lays the foundation for all methods discussed in the following section.

2.2.2 Methods

Since missing value problems are encountered in various disciplines, several review papers in different research areas have been published, such as statistics [185], operations management [299], psychology [162, 244, 253, 265, 116, 22], chemometrics [312, 313, 114]. Assuming data are missing at random (MAR) , this paper will summarize missing value methods that can be applied in the process industries and introduce the latest machine learning methods originated from computer science.

2.2.2.1 Deletion methods

The simplest procedure to handle incomplete data set is to eliminate any time point that contains missing values, known as list-wise deletion. However, such a procedure only works well for large data sets without significant changes, and it was pointed out that it will sacrifices a large amount of data, reduce the statistical power and lead to biased parameter estimation with more uncertainty [185].

Another deletion-based procedure is pairwise deletion, which only removes missing observations for certain variables at a time point if those are needed in further analyses, and the values of remaining variables at the same time stamp can still be used in other calculations. Monte Carlo simulation [162, 244] shows that pairwise deletion outperforms list-wise deletion in obtaining more accurate parameter estimation. However, because only missing values are deleted for certain variables, it will lead to inconsistency in time stamps for different variables, making it hard to reconstruct the original data set and interpret the pairwise correlation matrices [253].

2.2.2.2 Replacement methods

Instead of simply deleting missing elements in the original data set, the replacement-based procedures seek to impute them using available data and enjoy the advantages such as retaining the sample size and statistical power. However,such a procedure may introduce bias if the missing data are MNAR and correlations between variables are significant [185].

Generally, four types of replacement procedures can be used: mean-based, hot-deck, regression-based and interpolation-based.

- **Mean replacement**: We can substitute the missing items with either unconditional means (not dependent on other variables) or conditional ones. For categorical variables, instead of using means, we can use the mode (the value with the highest frequency).

- **Hot-deck replacement**: A missing value is filled with a value from anther similar case where the corresponding observation is complete. The hot-deck imputation can be divided into two stages: the data are first classified into separate and homogeneous clusters, and then the missing values are replaced with ones in complete cases at the same cluster. However, the hot-deck replacement method cannot be implemented if no similar case can be found.

- **Regression replacement**: The missing values in a certain variable ("target variable ") can be estimated based on other variables in the data set, assuming they are related to each other. A simple implementation is to fit a linear model for continuous observations (or logistic regression for binary variables) using other variables as predictors and the target variable as response based on available data, and estimate the missing elements in the target variable using calculated regression coefficients or weights [184]. However, it is cautioned that the predicted values do not necessary fall within logical limits of the target variable. Moreover,

such a procedure will impose extra factor structure on the original data set and we should avoid including independent variables in such a linear model [244].

- **Interpolation replacement**: For a dynamic data set, assuming the data are missing at random (MAR) and the correlation between variables are low, the interpolation procedure can help retain the dynamic nature of the target variable by fitting a polynomial based on neighboring points and use it to predict the missing points. However, these interpolation methods can not be applied alone, because sometimes there are not enough neighboring points to be used in polynomial fitting.

2.2.2.3 Model-based methods

The model-based methods assume that the observations of the target variable follows a distribution model and by estimating the model parameters based on observed data, we can infer the values of missing values. Generally, model-based methods include the multiple-imputation (MI) and two closely related routines: maximum-likelihood [9, 313]. Although the MI algorithm is less efficient than the maximum-likelihood(ML) algorithm, it accounts for the uncertainties caused by the missing values. For more information concerning MI, refer to [259], [264] , [185] and [22].

In semiconductor manufacturing processes, missing values are largely caused by inconsistent sampling strategies and outlier removal. A minimum norm estimation method incorporated with Tikhonov regularization algorithm

was proposed [235] to estimate the missing values for run-to-run EWMA-controlled processes, which outperforms other ad hoc techniques, such as deletion and mean substitution. It can be used in real-time to produce forecasts for future batches in a run-to-run scenario; however, the performance of the method heavily depends on the missing patterns: there is no universal formula applicable to all missing patterns and an irregular pattern might lead to a singular matrix hard to be inversed. A missing values-Patient Rule Induction Method (mPRIM) [173] was proposed which systematically processes the incomplete data set and optimizes the manufacturing process at the same time. Though the mPRIM algorithm can be applied to incomplete data sets with moderate missing rates, it assumes that all variables follow a multivariate normal distribution, which might not be satisfied.

2.2.2.4 Machine learning(ML) methods

- **Unsupervised learning – clustering**

  The Gaussian mixture model is a desirable candidate for dealing with processes with multiple operating conditions because the data no long follow a unimodal Gaussian distribution [333]. In order to handle missing values, A multi-time-slice dynamic Bayesian network with a mixture of the Gaussian output (MT-DBNMG) method was proposed and tested on a continuous stirred tank reactor system and the Tennessee Eastman process [340].

  A fuzzy similarity-based reconstruction method [23] was proposed for

nuclear power plant on-line monitoring, where the data set exhibits multidimensional time series features. The method replaces the missing values with a weighted sum of reference trajectories, and the weights are proportional to the fuzzy similarities between the reference segments and the target segment with missing values.

- **Supervised learning– classification & regression**

  A decision tree in data mining is a predictive model that can be used for either classification or regression tasks [251]. It can be directly used to predict missing attribute values, and outperforms the "autoclass"algorithm [55], as shown in the experimental results [174]. For example, the C4.5 algorithm [241, 240] was applied to improve the system identification of blast furnace ironmaking process from incomplete industrial data [336].

  To solve the problem of over-fitting and pruning faced by decision tree algorithms, the "random forest"was proposed [45] which randomly samples all trees in a forest with the same distribution to construct a predictive model. The random forest not only is more powerful in predicting missing values than the decision tree, but also incorporates two ways to replace missing values in the model training process: a fast way uses medians of available observations for non-categorical variables and modes for categorical ones, and a slow way iteratively constructs a forest using means (non-categorical) or modes (categorical) of available values

weighted by calculated proximities. Because a sampling step is involved in the random forests procedure, the computational cost increases significantly in comparison with the decision tree.

A new missing data imputation strategy based on estimating pairwise distances has been proposed [92, 93]. To calculate the pair-wise distance metrics, several machine learning algorithms can be applied, such as the k-nearest neighbor (kNN), the support vector machine (SVM) [70] and the Gaussian mixture model [246]. The method incorporates the EM algorithm to calculate missing values and shows excellent performance when the missing rate is high.

- **Dimensionality reduction**

  Another way to deal with missing value is to modify the algorithms themselves so that they can handle incomplete data set. For example, principal component analysis (PCA) and partial least squares (PLS) are common dimensionality reduction techniques employed to infer the latent structure in the data, and several procedures can be incorporated into the decomposition process and make them still applicable when dealing with incomplete data sets.

  – **Matrix factorization**: The decomposition of the original data set to find principal components is equal to minimizing the following

objective function:

$$\phi = \left\| \mathbf{Y} - \mathbf{T}\mathbf{P}^T \right\|^2 = \sum_{ij} \left(Y_{ij} - \sum_k t_{ik} p_{jk} \right)^2 \tag{2.4}$$

where $\mathbf{Y}$ is the complete data set, $\mathbf{T}$ is the score matrix and $\mathbf{P}$ represents the loading.

Usually, the alternative least squares (ALS)[107, 82] and singular value decomposition [118] can be applied to minimize the objective equation shown in Eq. (2.4). If the dimension or rank of matrix $\mathbf{Y}$ is low (often the case in computer vision and patter recognition field), the Wiberg method [318] can be used to estimate missing values in PCA modeling [278, 225, 94].

– **Other methods**:

A single component projection (SCP) derived from the NIPALS algorithm [68, 112] was proposed for PCA and PLS model building with missing data [219, 218], and it has been applied in concurrent PLS-based process monitoring with incomplete data sets from the Tennessee Eastman process [344].

The maximum likelihood was extended to the framework of PCA to handle missing data faced in multivariate data analysis [315], and it was later improved using the EM algorithm (EM-PCA) [313, 214, 42]. A similar method called the conditional mean replacement (CMR) or the known-data regression (KDR) algorithm

has been proposed [219] where only the expectation step of the EM algorithm is incorporated, and has been extended to dealing with missing data in the context of exploratory data analysis [17, 18, 50].

A method based on nonlinear programming (NLP) was proposed to estimate parameters of PCA in the presence of missing observations [81]. The NLP method bas been compared with the NIPALS-based one through a simulation example and a case study from pharmaceutical industry. An NLP-based method for PLS application was also reported [237].

Recently, Bayesian inference has gained popularity among the artificial intelligence and machine learning community. Methods combining the EM algorithm and the Bayesian approach can be used to solve missing value issues faced in PLS modeling [160], identifying nonlinear parameter varying systems [85], and state estimation of a batch process [342]. Other Bayesian methods include the data augmentation, which was used for sampling or iterative optimization by introducing latent variables [302, 142].

- **Neural network**

The multilayer perceptron(MLP) belongs to feed-forward artificial network models that map input data to outputs [252]. The MLP was

proposed to deal with incomplete data set containing categorical variables [279], and obtained a better result than classic procedures such as mean/mode-replacement, regression and hot-deck.

The self-organizing map (SOM) [165] and the generative topographic mapping (GTM) [36] are especially useful in inspecting and visualizing high-dimensional data. The GTM incorporated with the EM algorithm shows better performance in dealing with incomplete data sets [305].

2.2.3 Summary and implications

Table 2.1 provides a critical overview and assessment of each missing data imputation method covered in this section. As shown in the tables, there is no universally applicable missing imputation method. When selecting algorithms for a given incomplete data set, several basic factors have to be considered. First is whether there exists a similar but complete data set that can be used either as a reference to perform hot-deck or a training data set to train a model used for missing value prediction. Second is the dimension of the data set; we should balance between the computational cost and quality of the prediction. For example, although techniques using maximum-likelihood can provide the most probable value of missing observations, the process to obtain such a result will be much more complicated than a simple mean-imputation. Third is the fraction of missing observations α. If α is sufficiently

small (for example, $\alpha < 1\%$), a simple pre-processing step can already obtain desired results and any deviations introduced due to inaccurate imputation may be neglected. However, with a larger α, errors arising from the simple pre-processing step become increasingly relevant, and we have to not only examine the patterns or statistical properties of missing data more carefully, but also consider the model performance as well. A case study with different levels of missing data can be found in [189].

Table 2.1: Critical overview and assessment of missing value imputation methods

Methods	Selected authors	Advantages	Disadvantages
List-wise deletion	Little and Rubin (2002)	- Simple - Applicable to data sets with a low missing rate	- Sacrifice a large amount of data - Reduce statistical power - Bias parameter estimation
Pairwise-deletion	Raymond and Roberts (1987) Roth (1994) Kim and Curry (1977)	- Simple - Applicable to data sets with a low missing rate	- More accurate parameter estimation than list-wise - Generate inconsistency in time stamps of variables - Hard to reconstruct original data and interpret pairwise correlation matrix
Mean replacement	Little and Rubin (2002)	- Simple - Return a complete data set	- Deflate the variation of a variable - Bias the confidence interval
Hot-deck replacement	Cheeseman et al. (1988) Lakshminarayan et al. (1999)	- Take advantage of information from similar cases	- Hard to implement if no similar case exists - May fail to account for the uncertainty of missing values
Regression replacement	Raymond and Roberts (1987) Little (1988)	- Take advantage of relations between variables - Simple and interpretable	- Predicted values might surpass the limit - Do not apply on independent variables
Maximum-likelihood (ML)	Allison (2012) Hansson and Wallin (2012)	- Maximize the chance of finding the relations observed in the data - Has solid theoretical basis	- Hard to derive an analytical expression for a likelihood function - Not working when data do not follow a certain distribution
Expectation-maximization (EM)	Walczak and Massart (2001b) Gupta and Chen (2011) Zhou and Lim (2014)	- Similar to ML procedure, but no analytical expression for likelihood function is needed	- Iterative procedures with a low speed of convergence - Easily fall into local optimum - High computational cost
Multiple imputation (MI)	Rubin (1987) Schafer (1997) Baraldi and Enders (2010)	- A distribution is obtained for a single missing value - Account for the uncertainties of missing values	- Less efficient than ML or EM algorithm
Minimum norm	Prabhu et al. (2009)	- Performs well in case of varying sampling rates and gain mismatch - May be used in real-time to produce forecasts for future batches in a run-to-run scenario	- An irregular missing pattern might leads to a singular matrix hard to be inversed - Different formula have to be generated for different missing patterns
m-PRIM	Kwak and Kim (2012)	- Applicable to data sets with moderate missing rates - Simultaneously handles missing values and improve process	- Not efficient - Assumption that all variables follow multivariate normal distribution might not be satisfied

Table 2.1: Critical overview and assessment of missing value imputation methods (continued)

Methods	Selected authors	Advantages	Disadvantages
MT-DBNMG	Zhang and Dong (2014)	- Applicable to incomplete data set not following a unimodal Gaussian distribution - Robust to unknown noises - On-line implementation	- Needs a complete historical data to design and estimate MT-DBNMG
Fuzzy similarity based	Baraldi et al. (2014)	- On-line implementation - Applicable to time dependent data sets	- Complete reference trajectories might not be available - Difficult to set appropriate tuning parameters to calculate similarities and weights for different data sets
Decision trees	Quinlan(1986, 1993) Zeng and Gao (2009)	- Simple to implement and interpret - Can use both categorical and continuous values - Fast once rules are developed - Robust to outliers	- A training data set with an appropriate size is needed - Models are sensitive to small variations in data set and easily overfit - Pruning and simplifying the tree is needed
Random forests	Breiman(2001)	- Simple to implement and interpreted - Can contain both categorical and continuous values - Fast and scalable - Avoid over-fitting and pruning - Robust to outliers	- A training data set with an appropriate size is needed - A high computational cost
Pair-wise distance	Eirola et al. (2013, 2014)	- Applicable to data sets with a high missing rate - Account for the uncertainty in imputation	- Not effective when dealing with high dimensional data - High computational cost
Matrix factorization	Gabriel and Zamir (1979) De Ligny et al. (1981) Grung and Manne (1998) Okatani and Deguchi (2007) Walczak and Massart (2001a)	- Simple and interpretable - Take advantage of all information available	- An optimization step is involved which will slow down the process when dealing with a large data set - Predicted values are not necessarily within the limits of a certain variable - The covariance matrix might be distorted and affect model order selection
Single component projection (SCP)	Nelson et al. (1996) Nelson (2002) Zhao et al. (2014)	- Simple and interpretable - Take advantage of all information available	- Errors will increase with the collinearity of the loading vectors and noise variances - Errors will propagate from one score to another through deflation - Perform poorly when critical combinations of measurements are missing

Table 2.1: Critical overview and assessment of missing value imputation methods (continued)

Methods	Selected authors	Advantages	Disadvantages
Non-linear programming	De la Fuente et al. (2010) Puwakkatiya-Kankanamage et al. (2014)	- Guaranteed orthogonal loadings and scores - A faster convergence rate - Simple to interpret the model - Applicable to data sets with moderate missing rates	- Have to carefully selecting number of principal components - May get local optimum solution and complicate the problem - Model is sensitive to small data changes
ML-PCA/EM-PCA&PLS	Wentzell et al. (1997) Walczak and Massart (2001b) Serneels and Verdonck (2008)	- Maximize the chance of finding the relations observed in the data - Account for the uncertainties caused by missing values	- The covariance matrix may be singular - Not working when data do not follow a certain distribution and easily fall into local optimum
CMR (KDR)	Nelson et al. (1996) Arteaga and Ferrer (2002, 2005) Camacho (2010)	- Simple and interpretable - Take advantage of all information available	- Deflate the variation of a variable - Bias the confidence interval
EM-Bayesian	Khatibisepehr and Huang (2008) Deng and Huang (2012) Zhao et al. (2013a)	- Combine prior information with data within a solid theoretical framework - Maximize the chance of finding the relations observed in the data - Provide interpretable answers	- Not efficient - Hard to select an appropriate prior, which will not heavily influence the posterior distributions - A high computational cost - Data set might not follow a certain distribution which can be used for Bayesian inference
Bayesian data augmentation	van Dyk and Meng (2001) Imtiaz and Shah (2008)	- Prevent the deflation of the covariance matrix - Predict better than matrix-factorization methods - Withstand a high missing rate (up to 25%)	- Fluctuations are shown during the converging process - An iterative procedure increases the computational cost - Data set might not follow a certain distribution which can be used for Bayesian inference
Multiplayer perceptron (MLP)	Silva-Ramírez et al. (2011)	- Adaptive learning - Can contain both categorical and continuous values	- A training data set is needed - Has to carefully tune the learning rate - High computational cost
Self-organizing map (SOM)-based/ Generative topographic mapping (GTM)-based	Vatanen (2012) Vatanen et al. (2015)	- Easy to interpret and visualize - Suitable for high-dimensional data - Maximize the chance of finding the relations observed in the data	- Require careful initialization - Difficult to select the number of map units - Iterative procedures with a low speed of convergence - Easily fall into local optimum

2.3 Outlier detection

2.3.1 Motivation

Outliers are defined as observations or subsets of observations that do not show a consistent behavior with the rest of data set from a statistical perspective [26], and they usually have to be removed before conducting the data mining step, such as system identification, because the results of model parameter estimation and data analysis might be negatively affected. In the process industries, various reasons can cause outliers, such as malfunction of sensors and inappropriate treatment of missing data.

There are two types of outliers: univariate and multivariate. Univariate outliers occur only within a single variable context, while multivariate ones appear when combinations of variables violate a certain boundary. Usually, multivariate outliers are treated the same way as univariate ones for simplicity; however, such a procedure has several drawbacks [298]: (1) the interaction between variables might lead to over-specification of the number of outliers when neighbors of a single variable exhibit abnormalities; (2) dynamic changes of groups of variables might mask multivariate outliers so that the latter will not show significant deviations.

Another way of categorizing outliers is based on their impact on other observations, especially in a time-dependent data set: additive outliers (AOs) and innovational outliers (IOs). While the AOs affect the observation only at that exact time point, the IOs impact a finite number of samples in stationary processes, e.g., an auto-regressive moving average (ARMA) process, and exert

permanent effects on the following observations, leading to a transient level change (TC) or a level shift(LS) for non-stationary processes, e.g., an auto-regressive integrated moving average(ARIMA) process [56, 43]. In practice, while inappropriate treatment of missing data is likely to cause AOs, the IOs are generated due to process disturbances.

2.3.2 Terminology

2.3.2.1 Breakdown point (BDP)

The breakdown point was defined as the largest percentage of outliers that an algorithm could withstand [120]. We can view the influence of outliers from two opposite angles: the masking effect and swamping effect, which evaluate their impact on the method's detection rate and mis-identification rate (type I error) respectively. Consequently, the breakdown points can be categorized into masking and swamping breakdown points [76].

2.3.2.2 Outlier region

The definition of outlier region can be shown in the following example: for any $\alpha\,(0 < \alpha < 1)$, the α outlier region of the $N\left(\mu, \sigma^2\right)$ is defined by Eq. (2.5):

$$out\left(\alpha, \mu, \sigma^2\right) = \left\{x : |x - \mu| > z_{1-\alpha/2}\sigma\right\} \tag{2.5}$$

where z_q is the quantile function which can be calculated from the inverse of the cumulative distribution function(CDF), Φ, of the standard normal distribution

$N(0,1)$, i.e.:

$$z_q = \Phi^{-1}(q) = \sqrt{2} erf^{-1}(2q-1) \tag{2.6}$$

where $erf^{-1}(x)$ represents the inverse of error function.

For example, if $\alpha = 0.05, z_q = z_{1-\alpha/2} = z_{0.975} = 1.96$, which indicates that for sequence $\{x_k\}$ subject to $N(0,1)$, the outlier region is composed of the 5% of total samples with an absolute value larger than 1.96.

2.3.3 Methods

Numerous methods have been provided for outlier detection, and there are good review papers written by experts from different disciplines, such as computer science [132], chemometrics [280, 75], etc. Generally, based on whether we have knowledge on the process model a priori or not, we can categorize them into model-based methods and data-driven methods. In this paper, we further divide data-driven methods into univariate and multivariate based on the nature of outliers. While univariate methods are mainly statistical, the multivariate ones include more advanced machine learning procedures.

2.3.3.1 Data reconciliation and gross error detection methods

Data reconciliation is a technique developed to reduce the effects of random errors in the data and to obtain more accurate measurements as a result. The key feature of data reconciliation is that it makes use of process model constraints and makes sure all the measurements satisfy those constraints. There are two types of gross errors: one is measurement related such

as malfunctioning sensors and the other is process-related such as process leaks. A good review of data reconciliation and gross error detection methods for steady-state and linear dynamic systems can be found in [216], together with industrial applications and software available. Usually techniques from linear programming [181] or mix-integer programming [282] are incorporated in solving such a problem. For nonlinear dynamic systems, extended methods were discussed in [179], [206], [58], [59], [281], and [3].

2.3.3.2 Resistant regression methods

From the perspective of regression theory, removing outliers from data sets is equivalent to estimating the underlying model of the "meaningful" values. For data set $Z = \{(x_{11}, ..., x_{1p}, y_1), ..., (x_{n1}, ..., x_{np}, y_n)\}$, assume a linear model exists as shown in Eq. (2.7):

$$\hat{y}_i = x_{i1}\hat{\theta}_1 + ... + x_{ip}\hat{\theta}_p \tag{2.7}$$

where $\hat{\theta} = \left[\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p\right]^T$ are regression coefficients, $\{x_{i1}, x_{i2}, ..., x_{ip}\}$ are predictor values, $\hat{y}_i$ is the estimated value of y_i, and the residual is defined in Eq. (2.8):

$$e_i = y_i - \hat{y}_i \tag{2.8}$$

Several regression techniques, or known as estimators with intrinsic robustness when faced with outliers are summarized in Table 2.2. The reference method is ordinary least squares (OLS). Although the estimators are simple and take advantage of relations between variables, they do not apply to in-

dependent variables and the iterative procedures of deleting and refitting will significantly increase the computational cost.

Table 2.2: Robustness attributes of various regression estimators [16]

Estimator	Breakdown point(BDP)	Efficiency
OLS	0	100
LAV [87]	0	64
LMS [255]	0.5	37
LTS [255]	0.5	8
R-estimates [144]	< 0.2	90
M-estimates [137]	0	95
GM-estimates [198]	$1/(p+1)$	95
S-estimates [254]	0.5	33
GS-estimates [72]	0.5	67

2.3.3.3 Proximity-based methods

Another category of outlier removal methods, proximity-based methods, are based on estimating data location and scale that can be used to calculate the outlier region defined in the previous section. It is important to point out that a critical assumption of proximity-based methods is that the data are independently and identically distributed (i.i.d.) or follow a multivariate normal distribution. Such an assumption is usually compromised by process dynamics, and to deal with outlier detection in a dynamic data set, we can turn to time series methods discussed later. Commonly used proximity-based methods are introduced below.

- **Univariate outlier detection**

– 3σ **rule** The 3σ rule has been widely used for detecting outliers from an independent and identically distributed (i.i.d.) data set $\{x_k\}$ subject to a normal distribution $N(\mu, \sigma^2)$. If the following condition holds:

$$|x_k - \mu| > 3\sigma \tag{2.9}$$

then x_k is an outlier. A similar outlier detection procedure called the exteme studentized deviate (ESD) identifier was found in [76]. Instead using a constant value 3 in Eq. (2.9), the ESD-identifier employs a function $g(\alpha, N)$, which can be calculated based on the number of samples N and the significance level α.

– **Hampel identifier** Instead of using mean and standard deviation in Eq. (2.9), the Hampel identifier [121] uses the median med and median absolute deviation (MAD), as shown in Eq. (2.10):

$$|x_k - med| > g(\alpha, N)\, MAD \tag{2.10}$$

where $g(\alpha, N)$ is the function of aggressiveness, the med and MAD can be calculated by:

$$med = \frac{x_{[(N+1)/2]:N} + x_{[N/2]+1:N}}{2} \tag{2.11}$$

$$MAD = med\left(|x_1 - med|, ..., |x_N - med|\right) \tag{2.12}$$

where $[A]$ rounds A to the nearest integer less than or equal to A, and $x_{1:N}, ..., x_{N:N}$ are the ordered sequence of $\{x_k\}$.

– **Quartile-based identifier and boxplots** Unlike the 3σ rule or the Hampel identifier, the quartile-based detector uses the interquartile distance (IQD) Q as the scale parameter, which can be calculated by Eq. (2.13):

$$Q = Q_3 - Q_1 \tag{2.13}$$

where Q_1 is the lower quartile, $x_{0.25}$, and Q_3 is the upper quartile, $x_{0.75}$. Based on the definition of quartile, we can easily derive another expression for the median calculation shown in Eq. (2.14) in parallel with Eq. (2.11):

$$med = \frac{Q_1 + Q_3}{2} \tag{2.14}$$

Thus, for a symmetric data distribution, we can obtain the following condition to detect outliers [300]:

$$|x_k - med| > 2Q \tag{2.15}$$

A boxplot can be used as a graphical demonstration of the quartile-based detector, as shown in Fig. 2.2:

As shown in the figure, any point that lies outside the upper or lower fences, corresponding to Eq. (2.15), is considered as an outlier.

Other proximity-based methods include the Rousseeuw identifier [255], which uses the middle point as the location estimator and

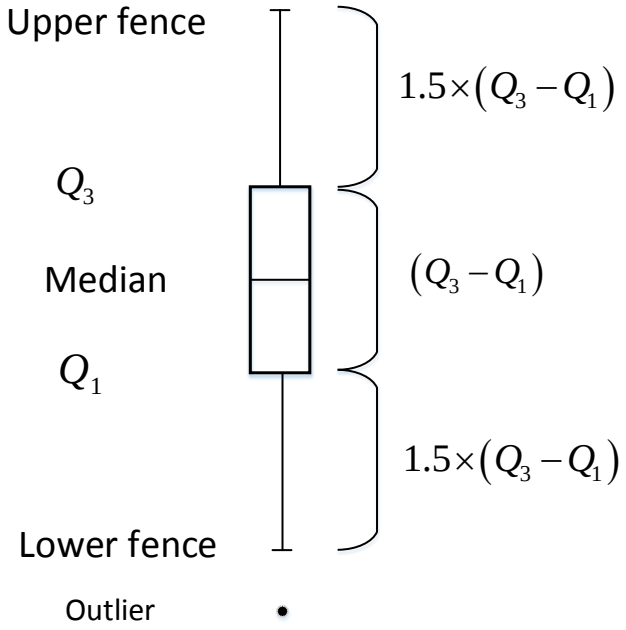


Figure 2.2: A typical boxplot [105]

the length of the shortest half sample size as the scale estimator, and the sequential extreme deviate removal-extreme studentized deviate(EDR-ESD) identifier [76]. A summary of breakdown points of proximity-based methods is shown in Table 2.3.

Table 2.3: Breakdown point (BDP) of outlier identifiers

Identifiers	Procedure	Swamping BDP	Masking BDP
3σ rule/ESD	single-step	1	1/(N+1)
Hampel	single-step	$\to 0.5, as\ N \to \infty$	0.5
Quartile-based	single-step	0.5	0.25
Rousseeuw	single-step	$\to 0.5, as\ N \to \infty$	0.5
EDR-ESD	sequential	0.5	$\min\left(\frac{1}{2}, \frac{k^*}{n+k^*}\right)(I)$

(I)$k^* = \min\left\{j \geq 1 : n + j - 1 < t^*_{n+j}(\alpha)^2\right\}$ or $k^* = \infty$ if no such j exists

- **Multivariate outlier detection**

  We can extend the 3σ rule to the multivariate case and replace μ and σ in Eq. (2.9) with the multivariate mean μ and the covariance matrix $\mathbf{\Sigma}$, as shown in Eq. (2.16):

$$\mu = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k,\ \mathbf{\Sigma} = \frac{1}{N-1}\left(\mathbf{x}_k - \mu\right)\left(\mathbf{x}_k - \mu\right)^T \tag{2.16}$$

  A minimum covariance determinant (MCD) estimator was proposed [255] based on Eq.(2.16), which tries to find a subset S^h of size $h\,(N/2 \leq h < N)$ of the original data set that minimizes the determinant of $\mathbf{\Sigma}^h$. Later a re-weighted strategy was incorporated to improve its robustness ($BDP = \frac{N-h}{N}$) and efficiency [258].

Because the covariance matrix $\mathbf{\Sigma}$ is widely used in multivariate data analysis, incorporating the MCD estimator enhances the capabilities of many methods to resist outliers, such as the Hotelling's T^2 used in PCA and related diagnostic plots [234, 322, 307], and canonical correlation analysis (CCA)[101].

2.3.3.4 Time series methods

In process industries, sometimes the "time" factor should be involved in the data analysis, especially for dynamic datasets. As mentioned in the previous section, some statistical detection methods [120, 121, 26, 255, 256, 231] no longer obtain a satisfactory performance because the assumption of i.i.d. is compromised. Traditionally, a moving window technique can be employed when dealing with time-varying datasets, because a key assumption is that the data can be treated as *i.i.d.* within a small enough window size. However, such a solution does not always give satisfactory results, especially when the dynamics is significant in the data set. Another solution is to first approximate the variations using time-series models, such as an autoregressive (AR) model, and then separate observations that show inconsistencies with the imposed models using outlier detection methods discussed in the time series literature, such as likelihood-based methods [103, 54, 56, 148]. In this section, common outlier detection techniques within the time series context are summarized in Table 2.4 with corresponding pros and cons. Although time series methods can process a dynamic data set within a solid theoretical framework and can

even detect outliers with different properties, it is hard to derive corresponding mathematical formulas and a high computational cost makes them only applicable to small data sets.

Table 2.4: Critical overview and assessment of time series methods

Methods	Models	Selected authors	Outlier type	Advantages	Disadvantages
Likelihood -based	AR	Fox (1972)	AO, IO	- Outperforms methods which are based on the i.i.d. assumption	- Only detects a single outlier in the data set
	ARIMA	Chang et al. (1988)	AO, IO	- Applicable to parameter estimation of ARIMA models in the presence of outliers	- Only one or two outliers are present in the data set
	AR/MA /IMA/A RMA	Chen and Liu (1993)	AO, IO, TC, LS	- Applicable to parameter estimation of different models in the presence of outliers	- At most three outliers are present in the data set - Level shifts might be misidentified as innovative outliers
	AR/ MA/ ARIMA	Jesús Sánchez and Peña (2003)	AO,IO, LS	- Prevents the confusion between IO and LS - Detects multiple outliers - Provides a joint procedure for estimating model parameters unbiasedly and detecting outliers	- A complicated procedure applicable only to small data sets with few outliers
	AR/ARI MA	Tsay (1988)	AO,IO, LS, TC	- Applicable to all types of outliers	- Difficult to tune
Deletion diagnostic based	ARIMA	Bianco et al. (2001)	AO, IO	- A robust filtering procedure reduces the effects of outliers - Works well when outlier contamination rate is as high as 10%	- An iterative procedure with a high computational cost, works well on small data set only
	ARMA	Abraham and Chuang (1989)	AO, IO	- Distinguishes an additive outlier from an innovational outlier - Includes an iterative deletion procedure to handle masking effects of outliers	- Works only on small data sets with only a few outliers - Hard to tune
Influence functional (IF) based	AR/MA	Martin and Yohai (1986)	AO, IO	- Applicable to both types of outliers - Solid theoretical framework	- Difficult to derive the influence functional - Only the AR (1) and MA (1) processes have been tested
	ARIMA	Peña (1990)	AO, IO	- Influence statistics based on the Mahalanobis distance takes into correlation of the data set	- Works only on small data sets with only a few outliers
Filter-cleaner	AR	Martin and Thomson (1982)	AO	- Simultaneously detects and replaces outliers	- A time series model is needed a priori
	ARMA/ ARIMA	Liu et al. (2004)	AO	- On-line implementation - Has a high break down point for ARMA process - Do not need a time series model	- Does not work well for ARIMA model - May inverse an ill-conditioned covariance matrix

Table 2.4: Critical overview and assessment of time series methods (continued)

Methods	Models	Selected authors	Outlier type	Advantages	Disadvantages
Likelihood-based	VAR	Franses and Lucas (1998)	AO, IO, LS, TC	- Applicable to all types of outliers - The identification and processing of outliers are both incorporated into the parameter estimation stage	- The diagnostic tool works well only for simple VAR models
	VAR	Lütkepohl et al. (2004)	LS	- Applicable to VAR processes with a structural shift at unknown time	- Works for data sets with only one simple level shift
	VAR/ VARIMA	Tsay et al. (2000)	AO, IO, LS, TC	- Based on both individual and joint likelihood ratio test statistics, which consider the information contained in single variables, as well as the interaction between variables	- Works only on small data sets with only a few outliers - Difficult to tune
Projection pursuit	VARIMA	Galeano et al. (2006)	AO, IO, TC, LS	- Detecting and processing multivariate outliers in some projection directions, which outperforms a direct testing on the time series	- Works only on small data sets with only a few outliers - Difficult to tune
Genetic algorithm (GA)	VARMA	Wiegand et al. (2009) Cucina et al. (2014)	AO	- Detect multiple isolated and consecutive additive outliers - Do not need a time series model	- An iterative procedure with a high computational cost, works well on small data set only
Time series Kalman filter (TSKF)	VARMA	Xu et al. (2015)	AO, IO	- No prior knowledge of the process model is needed - It is easy to tune - It can be applied to both univariate and multivariate outlier detection - It is applicable to both on-line and off-line operation	- Computationally intensive - Data sets with a high dimension will slow down the matrix calculation speed

2.3.3.5 Machine learning algorithm

- **Unsupervised learning – clustering**

  A multi-pass k-means clustering algorithm was proposed for novelty detection [8], which iteratively calculates every cluster's distance to the closest neighbor, compares it to a threshold, and decides whether to insert the story into the pre-defined k clusters or to create a new cluster, i.e., the novelty cluster. A similar approach is the k-medoids algorithm, which chooses a data point as the center instead of the mean value. Though the k-medoids algorithm is more robust to outliers than the k-means [41], it requires $O(n^2)$ running time whereas k-means is $O(n)$ [44]. The k-means or k-medoids algorithms obtain the best performance when clusters are distinct or separated from each other, and they generally have two disadvantages: one is that the number of cluster has to be specified a priori and the other is their inability to handle non-convex clusters of varying size and density.

  The hierarchical cluster analysis (HCA) technique was proposed to extract the hidden structure of objects through an iterative process that combines or separates object by object until all have been processed. One benefit of applying the HCA technique is that we do not need to specify the number of clusters. However, the HCA method is sensitive to outliers and fluctuations in the density of observations. A robustified HCA was proposed [10] which incorporates a self-consistent outlier reduction approach followed by the construction of a descriptive function,

but one must be careful in setting criteria for similarity/dissimilarity measurements because they significantly affect the results.

- **Supervised learning – classification and regression**

  - **Density-based methods**

    These types of methods involve a distance calculation such as Euclidean distance or Mahalanobis distance shown in Eqs. (2.17) and (2.18), which are used to estimate the local density (inversely proportional to distance):

$$D_{Euclidean} = \sqrt{\sum_{k=1}^{n} (x_i - y_i)^2} \tag{2.17}$$

$$D_{Mahalanobis} = \sqrt{(x - \mu)^T \mathbf{\Sigma} (x - \mu)} \tag{2.18}$$

    As shown in the above equations, the Mahalanobis distance is more computationally expensive than the Euclidean distance, especially for high-dimensional data. This is because it measures the distance from a point to the centroid μ scaled by covariance matrix $\mathbf{\Sigma}$, which requires the knowledge of the entire data set.

    From the perspective of distance, typical examples of the density-based methods are the kNN-based methods, such as retrospective kNN [164, 48], optimized kNN [243, 97], and weighted voting kNN [317].

From the perspective of density, by comparing the local density of an object to that of its neighbors, we can identify regions of similar density and define points with a lower value to be outliers. The concept of local outlier factor(LOF) was proposed [46] based on local density measurement to detect outliers, and extensions of the LOF can be found in [176], [169], [170], [277] and [194].

Although the density-based methods do not need to pre-specify the number of clusters and are applicable to clusters of arbitrary shape, there is a lack of theoretical rules for making outlier identification decisions, often the decisions are purely heuristic.

– **Support vector machine(SVM)**

The support vector machine(SVM) can be applied in outlier detection [289]. Support vectors are defined as data points which define the class boundary of normality. The general idea of SVM involves projecting input data onto kernels with a higher dimension; a kernel function is used to find a hyper-plane that serves as boundaries in between normal data and outliers. To mitigate the effects of outliers on model performance, an adaptive weighted least square support vector machine regression (AWLS-SVM) was proposed [74], which combines outlier detection approach and adaptive weight value for the training sample. Applying SVM-based methods will produce very accurate classifiers by solving a convex optimization problem and circumvent over-fitting; however, they require a long training

time and can only do binary classification.

- **Dimensionality reduction**

  The PCA method can be employed to detect outliers for high dimensional data sets by monitoring the Hotelling's T^2 and Q-statistics [168, 195, 98, 63]. However, the performance of PCA deteriorates when dealing with non-Gaussian distributed data set for two reasons. First, the mean and the variance-covariance of the process variables can no longer represent the characteristics of non-Gaussian data, which are commonly faced in process industries due to nonlinearities and the effects of feedback control. Second, both the Hotelling's T^2 and Q-statistics are developed based on a multivariate normal distribution assumption. Thus, satisfactory performance is unlikely to be obtained when the scores do not follow a normal distribution. Furthermore, even if the data follows Gaussian distribution, outliers will affect the location and scale estimation of the normal data. Thus, several robust PCA (RPCA) methods [57, 52] have been developed to enhance the robustness of the method.

  In addition, the assumption that the variables are uncorrelated in time in PCA no longer holds for dynamic processes with fast sampling [96]. To deal with a dynamic system, a dynamic PCA (DPCA) is proposed, which augments the data set composed of current values with past ones [172, 260].

  For outlier detection in processes where nonlinearities are present, ker-

nel PCA (KPCA) [268, 65, 66, 111] and independent component analysis (ICA) [153, 69, 141, 177] emerged as two solutions. In KPCA, first step is to perform a nonlinear mapping step that projects the input data into a linear feature space via a nonlinear function, and the next step is to implement PCA in the feature space. In ICA, the process data are decomposed into linear combinations of statistical independent components (ICs) that contain the main features of the process, during which higher-order statistics other than mean and variance-covariance are used. By monitoring the variability change in the dominant independent subspace, outliers can be detected.

To deal with outlier detection in a non-Gaussian distributed data set is to use a Gaussian mixture model(GMM) as shown in [34] and [249]. An extension of the GMM method is extreme value theory (EVT), which can be found in [248]. The EVT method uses a probability model to find outliers occurring in the tails of a distribution instead of applying a distance threshold like the GMM does.

For quality prediction using principal component regression (PCR) or partial least squares (PLS), a discussion of related outlier detection methods can be found in [203] and [232]. A methodology was later proposed to robustify the PLS method which involves dimension reduction, density-based clustering and outlier detection using a convex hull method [100, 99]. Although the robustified PLS method can directly improve the quality prediction results in the presence of outliers and simultaneously

detect outliers and inliers (points situated between clusters), it suffers from a relatively high computational cost and requires carefully selected splitting parameters as they significantly affect the outlier detection results.

- **Neural network**

  The multi-layer perceptron(MLP) with a single hidden layer was applied to detect outliers in jet engine data [215] and oil pipeline flow data [34] respectively. The convex hulls are defined by nodes in the first layer and used to detect outliers, and they outperform linear techniques in identifying the complex nonlinear class boundaries. An MLP with three hidden layers and three output and input neurons, also known as a replicator neural network (RNN) was proposed in [125], and it provided a measurement of the deviation of abnormal data records.

  An auto-associative (AA) neural network was used by [146] for outlier detection. The auto-associators are trained by normal data and used to reduce the number of hidden nodes and keep key attributes, analogous to PCA. The outputs of the auto-associators are fed into the network as inputs for reconstruction. By monitoring the reconstruction errors versus a certain empirical threshold, outliers can be detected.

  Self-organizing maps [165], which use vector quantization and nonlinear mapping to project the data distribution onto a lower dimensional grid network, can also be applied in outlier detection. After the new sample is

inserted into the network, the SOM algorithm compares it with existing data set, finds the best matching unit and updates the unit's as well as its neighbor's weighting vector. If the vector distance or quantization error between the best matching unit and the new sample is larger than pre-specified value, then an outlier is found [212]. A modified SOM – the habituating SOM (HSOM) was used in [202], and an iterative SOM approach with a robust distance estimation (ISOMRD) can be found in [49]. Both are more computationally efficient when addressing high dimensional data.

- **Other methods**

  A correntropy kernel learning (CKL) method was proposed for the identification of nonlinear systems with outliers and noise [187]. The correntropy (the name comes from correlation and entropy), which is an nonlinear similarity measure between two random variables, is used to evaluate the performance of identification models instead of traditional mean squared error criterion. The CKL method can reduce the effects of outliers by the use of a robust nonlinear estimator that maximizes correntropy. Another application of the maximum correntropy estimator in outlier removal can be found in [213], where a batch process data set is studied. Although the CKL method provides a brand new direction for system identification, more calculations are involved in such a method than in procedures designed based on the mean square error(MSE).

2.3.4 Summary and implications

Tables 2.5 summarizes the pros and cons of outlier detection methods covered in this section. As shown in the tables, there is no universally applicable outlier detection approach. When faced with various types of methods discussed in previous sections, we have to take into consideration numerous factors when selecting outlier detection methods such as the nature of outliers, robustness, and the capability to model the data distribution and separate the outliers from normal data. Although machine learning algorithms give promising results when handling complex cases, we cannot neglect the high computational cost, especially when the dimension of the data set is large. Last but not least, in many industrial cases, instead of preprocessing outliers, the process engineers are likely to remove them during the model building processes(PCA model, for example) as shown in Table 2.5.

Table 2.5: Critical overview and assessment of outlier detection methods

Methods	Selected authors	Advantages	Disadvantages
Data reconciliation and gross error detection	Liebman (1991) McBrayer and Edgar (1995) Chen et al. (1998) Narasimhan and Jordache (1999)	- Take advantage of process model constraints and make sure they are not violated - Reduce the random errors in the data	- Process model is needed - Original values will likely to be changed during a reconciliation process
Resistant regression methods	Anderson (2008)	- Simple and easy to interpret - Take advantage of relations between variables	- Iteratively deleting and refitting will increase the computational cost - Do not apply on independent variables
Proximity-based methods	Davies and Gather (1993) Rousseeuw and Leroy (1996) Becker (2000)	- Simple and easy to interpret	- The location and scale estimation will be affected by the outliers - The computational cost is high for multivariate outlier detection - The assumption that the data are i.i.d. or follow a multivariate normal distribution might not be satisfied
Time series methods	Fox (1972) Chang et al. (1988) Chen and Liu (1993) Bianco et al. (2001) Liu et al. (2004)	- Can handle dynamic data which are not identically and independently distributed - Have solid theoretical basis - Can detect outliers with different properties	- Hard to derive a mathematical formula - Computational cost is high - Only applicable to small data sets
K-means/K-medoids	Allan et al. (1998) Bradley et al. (1999) Bolton and Hand (2001)	- Fast and easy to interpret - Perform the best when clusters are distinct or well separated from each other	- Have to specify the number of clusters a priori - Work poorly when there exists an overlapping - Do not apply on categorical data - The centers of clusters have to be carefully initialized - Unable to deal with non-convex clusters of varying size and density
Hierarchical cluster analysis (HCA)	Almeida et al. (2007)	- Deterministic structure easy to read - Do not need to pre-specify the number of clusters	- Sensitive to the density of observations - Different similarity/ dissimilarity measurements may generate different results, and related criteria have to be carefully set - Less efficient
Density-based methods	Wettschereck (1994) Ester et al. (1996) Knorr and Ng (1998) Byers and Raftery (1998) Breunig et al. (2000)	- Discover clusters of arbitrary shape - Resistant to noise - Do not need to pre-specify the number of clusters	- Lack of clear rules for making outlier identification decision - Relatively high computational cost - High sensitivity to the input parameter setting

Table 2.5: Critical overview and assessment of outlier detection methods (continued)

Methods	Selected authors	Advantages	Disadvantages
Support vector machine (SVM)	Tax and Duin (2004) Cui and Yan (2009)	- Produce very accurate classifiers by solving a convex optimization problem - Less over-fitting and robust to noises	- Long training time, computationally expensive - Difficult to design and interpret the weights - Can only do binary classification
PCA-based methods	Kourti and MacGregor (1995) Ku et al. (1995) Chen et al. (1996) Russell et al. (2000) Chiang et al. (2003) Choi et al. (2005) Ge et al. (2009) Lee et al. (2011)	- Reduce the high dimensionality of the data set - Low computational cost - Can be robustified and modified to deal with data sets containing dynamics and nonlinearities	- Cannot obtain a satisfactory results when data do not follow a Gaussian distribution
Mixture models	Bishop (1994) Roberts and Tarassenko (1994) Roberts (1999) Zhu et al. (2014)	- Can obtain a satisfactory results when data do not follow a Gaussian distribution - The model training processes can be robustified	- Have to specify the number of mixtures a priori - Fail to work if the dimensionality of the problem is too high
PLS/PCR-related	Martens and Næs (1989) Pell (2000) Fernández-Pierna et al. (2002, 2003)	- Provide a visual aid to outlier detection - Simultaneously detect outliers and inliers (points situated between clusters) - Directly improve the quality prediction results in the presence of outliers	- Relatively high computational cost - Splitting parameters have to be carefully set because how the data set is split significantly affects the number of outliers
MLP/RNN/ AA neural network	Japkowicz et al. (1995) Nairac et al. (1999) Hawkins et al. (2002)	- Adaptive learning - Convex hulls defined by nodes outperform linear techniques	- A training data set is needed - The detection threshold has to be carefully set - High computational cost - Obtain a poor performance on sparse data set
SOM-based methods	Muñoz and Muruzábal (1998) Kohonen (1999) Marsland (2001) Yan (2011) Cai et al. (2013)	- Easy to interpret and visualize - Suitable for high-dimensional data	- Require a careful initialization - Difficult to select the number of map units - Iterative procedure which has a low speed of convergence
Correntropy kernel learning (CKL)	Munoz and Chen (2012) Liu and Chen (2014)	- Robust, applicable to identification of nonlinear systems with outliers and noises - Easy to initialize and set weights	- The kernel size has to be carefully selected - It needs more calculations than the mean square error (MSE) procedure

2.4 Noise removal and frequency analysis

2.4.1 Motivation

In process industries, most data come from sensors generating signals through electronic, electro-mechanical or electro-optical means. The sensor data are likely to be degraded by the environment, i.e., the signals are contaminated with high frequency noise. Sometimes, the noise must be inspected carefully because it may reflect operating condition changes, and an interdisciplinary approach called "acoustic chemometrics "is dedicated to exploiting noise to uncover information on the process [95]. The signal to noise ratio (SNR) is usually used to compare the magnitude of a desired signal to the level of background noise, and it is defined as the ratio of signal power to the noise power, often expressed in decibels, as shown in Eq. (2.19):

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{2.19}$$

where P is average power measured at the same or equivalent points in a system within the same bandwidth.

If the SNR is small, there are too many defects in the signal that can produce misleading or biased results in subsequent stages of data analysis and uses, e.g., controller optimization [209, 272], dead-time compensation [262] and system identification [60, 30]. Thus, it is necessary to extract useful information from measurement data through filtering.

2.4.2 Methods

Numerous techniques have been proposed to perform the noise filtering task, and they can also be categorized into two groups: model-based and data-driven. The most widely used model-based filter is the Kalman filter [155], and data-driven methods include digital filters [242, 190], Savitzky-Golay filter [263], wavelet filters [308], etc.

2.4.2.1 Model-based methods

Kalman filter estimates the hidden state of a linear discretized dynamic system from noisy observations through an iterative prediction-correction process. The process calculates the true values of states using incoming measurements and a process model, which is similar to the recursive Bayesian estimation procedure. In the process industries, most systems are nonlinear; thus, an extended Kalman filter (EKF) is often applied, which incorporates an extra step to linearize the model. Demonstrations of the EKF for chemical process systems can be found in [347], [157], [30], and [236]. More information concerning theories and applications of the Kalman filter can be found in [15], [126], and [47].

Particle filters are a type of filters based on a recursive Bayesian estimation procedure, which generate a set of samples, or particles via the Monte Carlo method, to approximate the posterior distribution. Similar to the Kalman filter, the objective of a particle filter is to estimate the posterior distribution of the hidden states from observed noisy measurements, but

it differs from the Kalman filter in that the process model can be nonlinear and there is no extra condition on initial states [62]. Particle filters have been applied to a number of industrial processes, such as tracking biological cells [276], a beer fermentation process [343], a polymerization process [61, 341], and a benchmark PH neutralization process [60]. A good review of theoretical development of the particle filters and their applications can be found in [62] and [227] .

A recursive multichannel instrumental variable (IV) lattice filter was proposed for a multivariate process contaminated with noise [180]. The lattice filter generates a residual vector to adaptively monitor multivariate dynamic processes. The filter has the capability to recursively update the time and order parameters of the process model and provides an monitoring index based on the Hotelling's T^2 statistic.

2.4.2.2 Data-driven methods

Sometimes we may not have a prior knowledge of the process model except for a desired or target process. Under such circumstances, adaptive filters, such as the least mean squares filter [128] and Wiener filter [321], can provide a linear time-invariant filtering of an observed noisy process by minimizing the mean square error (MSE) between the estimation and target signals. For more information about adaptive filtering, see [127].

If there is no target signal available, we can still apply other data-driven filtering methods, which clean the data by frequency thresholding. There are

general two types of digital filters: one is a finite impulse response (FIR) filter, the impulse response of which will only last for a finite time horizon N and finally settles to zero; the other is the infinite impulse response (IIR) filter, which will respond to the input infinitely. A thorough discussion on frequency-thresholding digital filter design implementation in MATLAB can be found in [190].

When performing tasks such as spectral analysis in quality control, common low-pass digital filters, such as exponential filters [314], no longer obtain a satisfactory result because the filter will erase any information in the high frequency domain, whether it is useful or not [228]. Moreover, most digital filters approximate current values based on previous observations. Extra time delays will be generated as a consequence. The Savitzky-Golay FIR low-pass filter [263], also known as polynomial least-square smoothing filter, outperforms common low-pass filters in preserving useful high-frequency information and prevention of extra delays. The Savitzky-Golay filter is described in Eqs. (2.20) and (2.21):

For a data sequence $\{x_t\}$, at one time point $t = 0$, we can obtain $2M+1$ samples centered at t, which can be used to fit the following k-th order polynomial with a minimized residual:

$$P_t = \sum_{i=0}^{k} a_k z^k; -M \leq z \leq M \tag{2.20}$$

where the coefficients a_k are calculated through minimizing the following ap-

proximation error:

$$E_m = \sum_{t=-M}^{M} (P_t - x_t)^2 = \sum_{t=-M}^{M} \left(\sum_{i=0}^{k} a_k z^k - x_t \right)^2 \tag{2.21}$$

A comparison of the commonly used first-order exponential filter (expressed in Eq. (2.22)) and the Savitzky-Golay filter is shown in Fig. 2.3: the exponential filter has a side effect of generating time delays which can be circumvented when using the Savitzky-Golay filter.

$$y_t = \theta x_t + (1 - \theta)\, y_{t-1} \tag{2.22}$$

where y_t is the filtered value at time t, x_t is the raw measurement and θ is the filter tuning parameter.

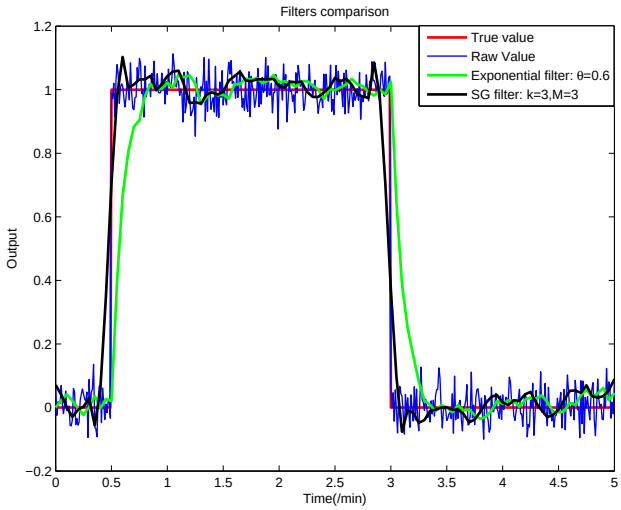


Figure 2.3: A comparison of first-order exponential filter and the Savitzky-Golay filter

Wavelet filters have been widely used in the fields of applied mathematics, signal processing and computer vision [308], and a few applications in

handling the process data have also been reported [21, 89, 109]. The wavelet filters generally remove noise in three steps: (1) decompose the signal into coefficients at several scales corresponding to different frequency levels by wavelet transform; (2) threshold the wavelet coefficients that correspond to undesired frequency components; (3) reconstruct the signal. Although the process is more complicated than in a digital filter and the Savitzky-Golay filter, it can be extended to multivariate de-noising and no extra time delays will be generated after finishing those steps [21, 13]. Moreover, incorporation of multi-resolution analysis (MSA) using wavelet transform provides us more valuable information on frequency change in time than the traditional frequency analysis using a Fourier transform, as shown in the next section.

2.4.3 Frequency analysis & multi-resolution analysis

The Fourier transform has been widely used in representing a function in its frequency domain, and it is defined in the formula shown in Eq. (2.23):

$$F(f) = \int_{-\infty}^{\infty} f(t)\, e^{-2\pi i f t} dt \tag{2.23}$$

As shown in above equation, the function $e^{-2\pi i f t}$ carries frequency information through parameter f, and the time information represented by t has been integrated out; as a consequence, we can no longer obtain information related to when a specific frequency component occurs or how different frequencies change with time. A solution is to use the wavelet transform, and the mother

wavelet function $\psi(t)$ is defined as:

$$\psi(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2.24}$$

As we can see in Eq. (2.24), the wavelet function differs from the frequency function $e^{-2\pi ift}$ used in Eq. (2.23) because it includes two parameters: a is the scale parameter preserving the frequency information, and b is the shift parameter containing time-domain information [197]. Similar to the Fourier transform, the continuous wavelet transform (CWT) of $f(t)$ is shown in Eq. (2.25):

$$W_{f(t)}(a,b) = \int_{-\infty}^{\infty} f(t)\,\psi^*_{a,b}(t)\,dt \tag{2.25}$$

An example demonstrating the superiority of multi-resolution analysis using wavelet transform is given in [4]. Suppose data follow a simulated model shown in Eq. (2.26):

$$y_t = \cos\left(\frac{2\pi}{p_1}t\right) + \cos\left(\frac{2\pi}{p_2}t\right) + \varepsilon_t; t = \frac{1}{10}, \frac{2}{10}, ..., 60 \tag{2.26}$$

where $p_1 = 12, p_2 = \begin{cases} 5 & ; 24 \leq t \leq 36 \\ 2 & ; otherwise \end{cases}$, and ε_t is white noise, $\varepsilon_t \sim N\left(0, 0.5^2\right)$. The signal contains mixed information at different frequency levels and goes through a temporary frequency change between $24 \leq t \leq 36$.

As depicted in Fig. 2.4, the Fourier transform shown in (d) fails to find the time when a certain frequency of the system changes, while the wavelet transform shown (b) preserves relevant information. The wavelet power spectrum in (b) can still be integrated over time, which obtains a similar result (c) as the Fourier transform result (d).

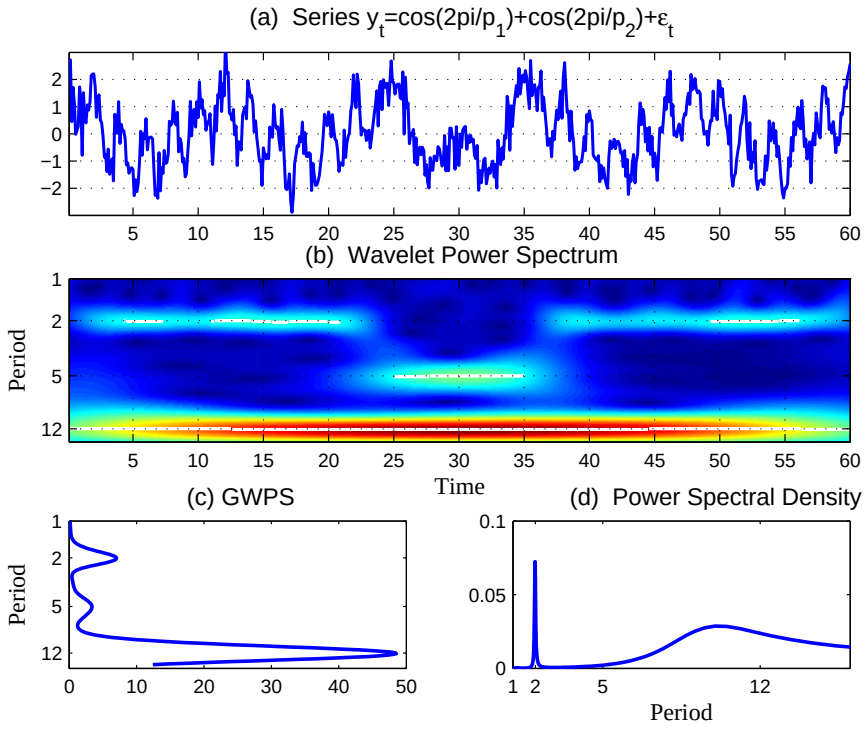


Figure 2.4: Wavelet transform demonstration: (a) Original signal. (b) Wavelet power spectrum of y_t, the color code for power ranges from blue (low power) to red (high power). (c) Global wavelet power spectrum- average wavelet power for each frequency. (d) Fourier power spectral density.

2.4.4 Implications

Noisy measurements should go through a certain filtering system, either model-based or data driven, to improve the data quality. To maximally preserve useful information and erase the effects of noise, we have to carefully tune the filters. Moreover, possible time delays caused by some filters should be taken into account. Nevertheless, frequency analysis and multi-resolution analysis are useful tools in information extraction, frequency selection and filter tuning.

2.5 Time delay estimation

Time delay estimation problems are frequently encountered in the process industries particularly due to flow transport or instrumental analysis, and usually input changes will not affect the quality parameters until after a certain period. On the one hand, because common linear models such as the PLS model do not take into account time delays, it is necessary to align the inputs with the outputs based on input-output time delays so that a better model prediction result will be obtained. On the other hand, input-output time delays provide process engineers with important information based on which control decisions can be made ahead of time. In addition, the performance of control synthesis techniques such as the minimum variance controller will be poor if the delay is not known a priori. A large time delay also limits performance of a PID controller, and negatively affects the closed-loop stability by causing phase lags. In those cases, the Smith predictor may be applied to do time-delay compensation [5, 271] and applying model predictive control (MPC) is an alternative solution.

A variety of TDE methods have been reviewed [37, 247, 339, 5], and they can generally be divided into model independent and model dependent methods [221]. The model dependent methods estimate the time delay as a model parameter or determine the time delay during iterative calculation [332, 346, 245]. A Laguerre transformation can be included in the methods to facilitate parameter estimation [102, 143]. However, some of the model dependent methods, such as the recursive least squares, might not be computation-

ally efficient, especially when the time delay is unknown. Thus, estimating the time delay by model independent methods a priori may drastically reduce the efforts for model parameter calculation and facilitate the system identification. Model independent methods are data-based, including the general cross-correlation method [163], the minimum entropy method [32], etc.

In the formulation of general cross-correlation, the cross-correlation function between two signals $y_1(t), y_2(t)$ is calculated by:

$$\hat{R}_{y_1 y_2}(\theta) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y_1(t)\, y_2(t-\theta)\, dt \tag{2.27}$$

The above cross-correlation result can be quickly obtained by applying Fourier transform and inverse Fourier transform, because it is proportional to the cross power spectral density function in the frequency domain, shown in Eq. (2.28):

$$\hat{R}_{y_1 y_2}(\theta) \propto \int_{-\infty}^{\infty} y_1^*(j\omega)\, y_2(j\omega) e^{-j\omega\theta} d\omega = \int_{-\infty}^{\infty} S_{y_1 y_2}(j\omega) e^{-j\omega\theta} d\omega \tag{2.28}$$

The time delay T_d is obtained by finding θ that maximizes the cross-correlation function shown in Eq. (2.28)

2.6 Summary

Heavily instrumented processes are challenged by the poor quality of a large amount of raw data and how to extract useful information from them. Data cleaning can improve the overall data quality for further analysis and model building, and it includes four components: missing data imputation,

outlier detection, noise removal, and time alignment. A method to clean the data is suitable in practice if it can:

- Incorporate the knowledge of the process and the requirements; for example, whether the process model is known
- Match the properties of the data set, such as dimension and the fraction of missing values
- Identify the mechanisms leading to contamination, such as the nature of outliers
- Keep the computational cost and tuning efforts within a manageable range when dealing with a large data set
- Preserve and reveal the nature or characteristics of normal data, such as the distribution or frequency

Because each process data set has its own distinguishing features, it is hard to generate a universally applicable data cleaning procedure. Based on the literature reviewed, we recommend the following directions for future research:

- Combining the data cleaning step with a later stage of model building and performance evaluation. By setting a certain threshold for model evaluation, such as R^2, we can adaptively tune the data cleaning step to prevent over-cleaning and optimize the model quality.

- On-line implementation of data cleaning methods. Most procedures discussed in this literature review are performed off-line. On-line data cleaning can provide a rapid methodology to prevent possible process deviations, but it poses challenges for the method's robustness, stability, and computational cost.

Chapter 3

Model impact analysis

Before studying on advanced and effective data cleaning methodology, it is necessary to investigate the impact of contaminated data on model performance and how data cleaning methods can improve it as a preliminary step. Such a step will provide valuable guidance in developing a new technique that combines the data cleaning step with model building and performance evaluation, as one of the research directions pointed out in Chapter 2.

3.1 Impact of outliers and noise on dynamic model identification

In this section, data from an industrial process will be used to demonstrate that outliers and noise negatively affect the process model identification and that data cleaning methods are useful in solving such problems.

3.1.1 Problem formulation

A non-linear CSTR with a cooling jacket process is modeled by Dayal and MacGregor [78]: At a constant volume, the reactant mass balance can be

expressed in Eq. (3.1):

$$V\frac{dc_A}{dt} = F_{in}\left(c_{A,in} - c_A\right) - Vk_0e^{-E/RT}c_A \tag{3.1}$$

The reactor energy balance is:

$$\rho_s c_{ps} V\frac{dT}{dt} = \rho_s c_{ps} F_{in}\left(T_{in} - T\right) + V\left(-\Delta H\right)k_0e^{-E/RT}c_A + UA\left(T_{cj} - T\right) \tag{3.2}$$

The cooling jacket energy balance is:

$$\rho_s c_{ps} V\frac{dT}{dt} = \rho_s c_{ps} F_{in}\left(T_{in} - T\right) + V\left(-\Delta H\right)k_0e^{-E/RT}c_A + UA\left(T_{cj} - T\right) \tag{3.3}$$

The conversion is defined by:

$$conversion = \frac{c_{A,in} - c_A}{c_{A,in}} \tag{3.4}$$

The parameters are shown in Table 3.1 [78, 250]:

The reactor feed rate F_{in} and cooling water flow F_w were used to control the conversion $c_{A,in}$ and temperature T respectively. PI controllers were used in the system and PRBS type perturbation signals were added to their outputs (as shown in Fig. 3.1). For more information concerning the setting of controllers, steady-states and so on, see the MATLAB simulink model built by Roffel and Betlem [250].

Both open-loop and closed-loop system identification (as shown in Figure 3.2) were conducted based on three different data sets obtained by following scenarios:

1. The raw clean data set obtained by simulating the reactor model, as shown in (a) $\sim$ (d) in Fig. 3.3.

Table 3.1: Specification of reactor parameters

Parameter	Physical meaning	Unit	Value
V	reactor volume	m^3	1.0
F_{in}	reactor feed	m^3/s	0.014
$c_{A,in}$	inlet concentration A	kg/m^3	866.0
k_0	pre-exponential constant	s^{-1}	$4 * 10^8$
E	activation energy	J/mol	$6 * 10^4$
R	gas constant	$J/\left(mol \cdot K\right)$	8.314
ρ_s	reactant density	kg/m^3	866.0
c_{ps}	specific heat of reactant	$J/\left(kg \cdot K\right)$	1.791
T_{in}	feed temperature	K	293.0
ΔH	heat of reaction	J/kg	-140.0
UA	product of heat transfer area and transfer coefficient	W/K	150
ρ_w	water density	kg/m^3	998.0
c_{pw}	specific heat of water	$J/\left(kg \cdot K\right)$	4.184
V_{cj}	volume cooling jacket	m^3	0.2
$T_{w,in}$	cooling water inlet temperature	K	290.0
F_w	cooling water flow	kg/s	18.161
T	reactor temperature	K	305
T_{cj}	cooling jacket temperature	K	300
c_A	concentration of A in the reactor	kg/m^3	346.4

2. The contaminated data set obtained by adding outliers ($|Amplitude| = 0.3$ *and* 3 for conversion and temperature) and zero-mean Gaussian noises($\sigma = 0.001$ *and* 0.1 for conversion and temperature) on the raw clean outputs to simulate sensor malfunctions and noisy environment, as shown in (e) $\sim$ (f) in Fig. 3.3.

3. The processed data set obtained by implementing 3σ rule (moving window size=20) and the Savitzky-Golay filter (k=3, M=5 and 7 for conversion and temperature) on the contaminated outputs.

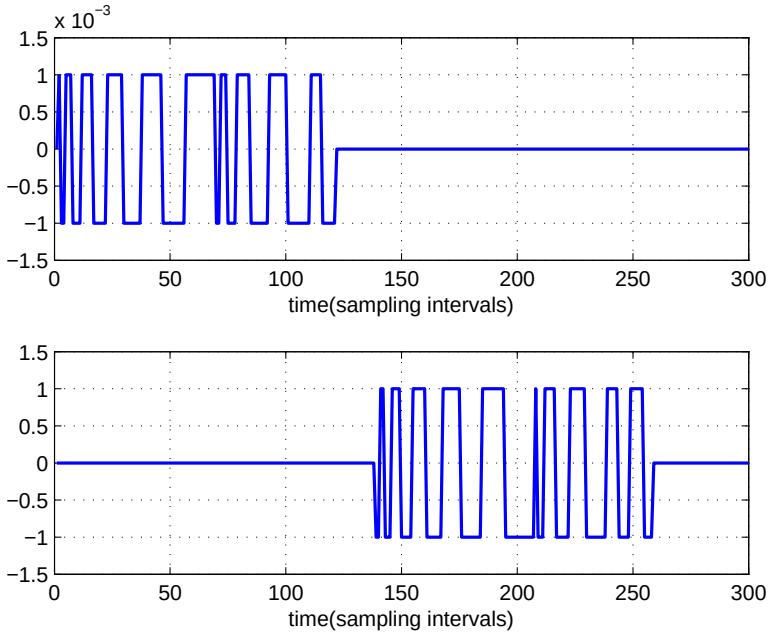
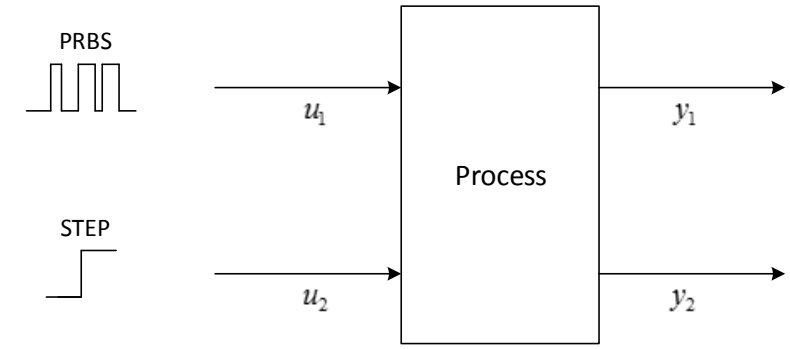
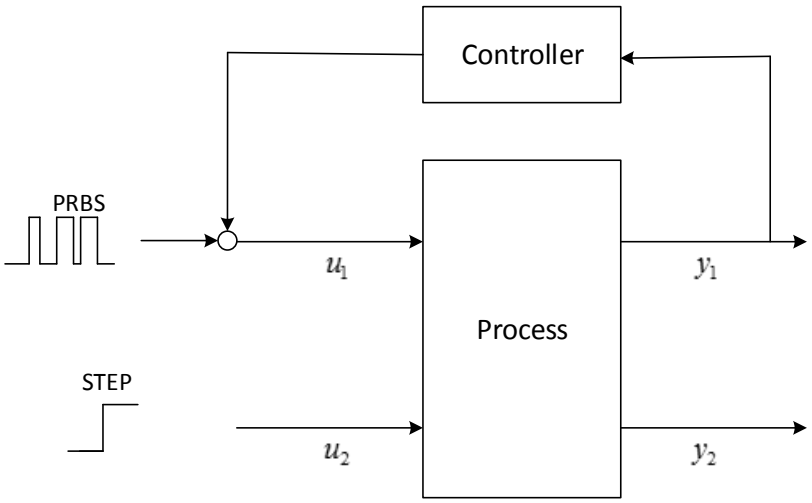


Figure 3.1: PRBS signals for F_{in} (top) and F_w (bottom) respectively

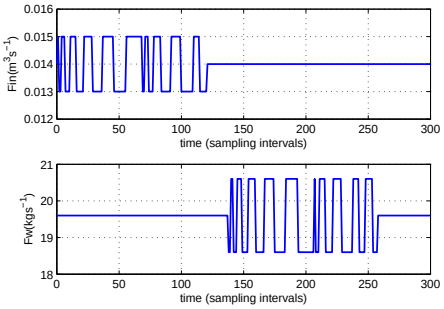


(a) Open loop

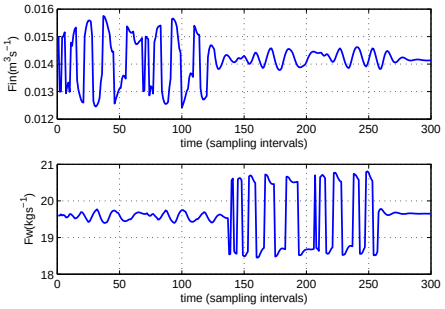


(b) Closed loop

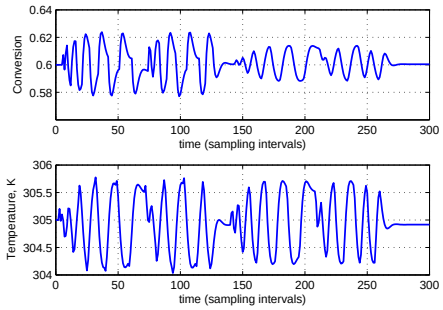
Figure 3.2: Types of experiments [250]



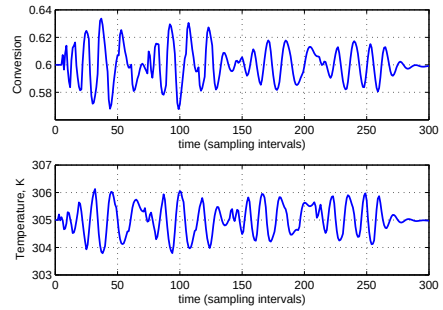
(a) Process inputs: open loop



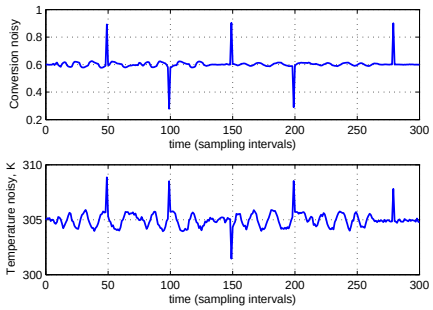
(b) Process inputs:closed loop



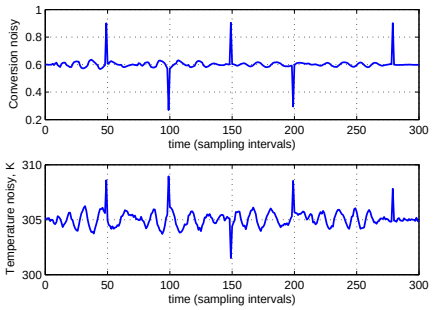
(c) Raw process outputs: open loop



(d) Raw process outputs:closed loop



(e) Contaminated process outputs: open loop



(f) Contaminated process outputs: closed loop

Figure 3.3: Process inputs and outputs

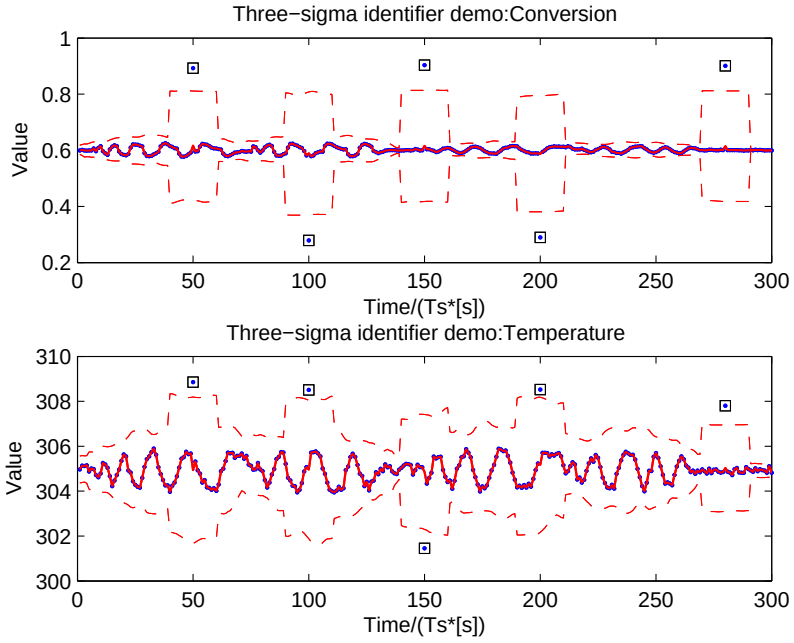
Table 3.2: Comparison of steady-state gains for open and closed loop identification: conversion

	$\delta C/\delta F_{in}$			$\delta C/\delta F_{cw}$		
	I	II	III	I	II	III
Calculated from model linearization in operating point	-5.57	-	-	-0.012	-	-
Closed-loop parametric model identification	-6.65	-4.99	-6.69	-0.011	-0.04	-0.01
Open-loop parametric model identification	-6.84	-13.25	-6.12	-0.011	-0.04	-0.01

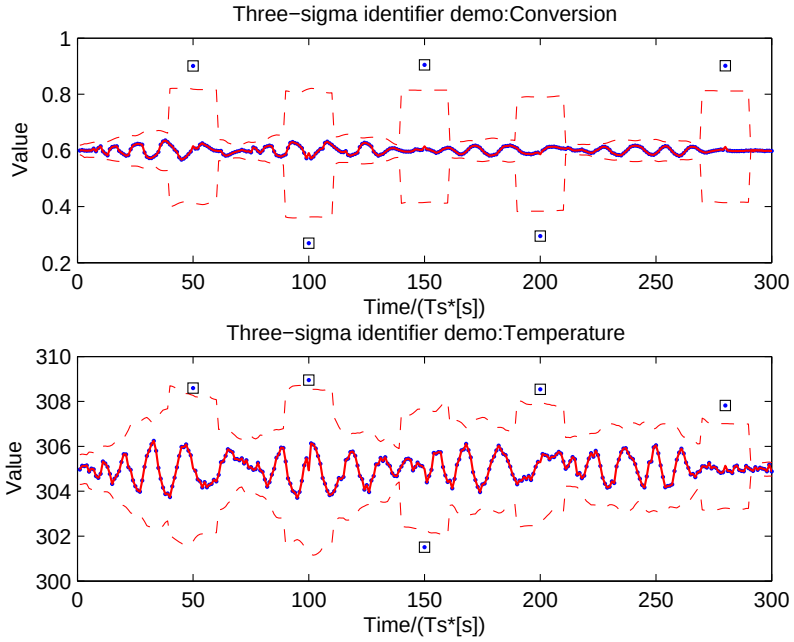
(I) Raw clean data
(II) Outlier contaminated noisy data
(III) Processed data

3.1.2 Results and discussion

Fig. 3.4 shows the outlier detection results of the process outputs: the 3σ rule successfully detects all the outliers. The resulting steady state gains were compared with that obtained from model linearization in operating point, as shown in Tables 3.2 and 3.3. While the open-& closed-loop results of scenario (I) significant differ from those of (II), they are quite close to (III). Moreover, in comparison with the model linearization results, the open- and closed-loop results of scenarios (I) and (III) make more sense than those of scenario(II). Thus, we can see that outliers and noises can significantly negatively affect the system identification, and it is necessary to process the data before conducting model identification.



(a) Open loop



(b) Closed loop

Figure 3.4: Outlier detection results

Table 3.3: Comparison of steady-state gains for open and closed loop identification: temperature

	$\delta T/\delta F_{in}$			$\delta T/\delta F_{cw}$		
	I	II	III	I	II	III
Calculated from model linearization in operating point	614.41	-	-	-0.63	-	-
Closed-loop parametric model identification	648.5	740.4	673.3	-0.50	-0.68	-0.50
Open-loop parametric model identification	630.2	743.7	640.1	-0.50	-0.65	-0.48

(I) Raw clean data
(II) Outlier contaminated noisy data
(III) Processed data

3.2 Impact of time delays on partial least square(PLS) models

The following simulated example demonstrate how time delays may affect the PLS model prediction in process industries.

3.2.1 Problem formulation

Assuming the quality parameter y is related to four inputs determined by Eq. (3.5):

$$Y(s) = \begin{bmatrix} 6 & -3e^{-3s} & \frac{1}{s+1}e^{-4s} & \frac{1}{2s+1}e^{-5s} \end{bmatrix} \begin{bmatrix} u_1(s) \\ u_2(s) \\ u_3(s) \\ u_4(s) \end{bmatrix} \tag{3.5}$$

where $u_1 \sim u_4$ are randomly generated binary signals, as shown in Fig. 3.5. PLS models were built based on two scenarios: one ignores the time delays while the other one does not and shifts the inputs correspondingly.

3.2.2 Results and discussion

The results shown in Fig. 3.6 demonstrate that the second scenario has a better model performance ($R^2 = 0.78$) than the first one ($R^2 = 0.44$) and that it is important to conduct the input-output time delay estimation before building models.

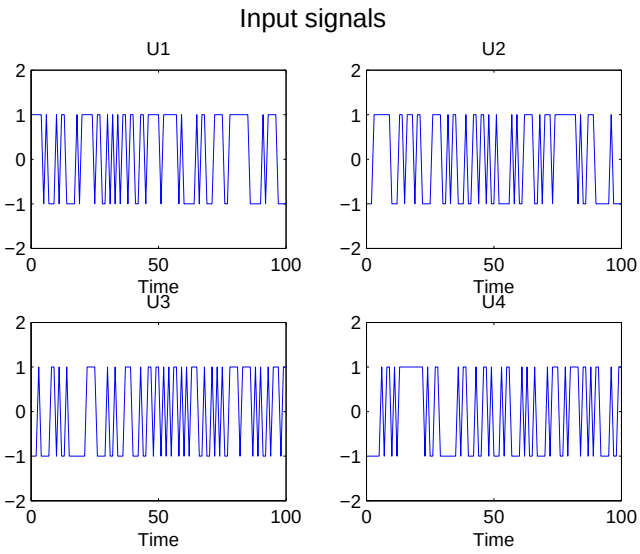


Figure 3.5: Process inputs

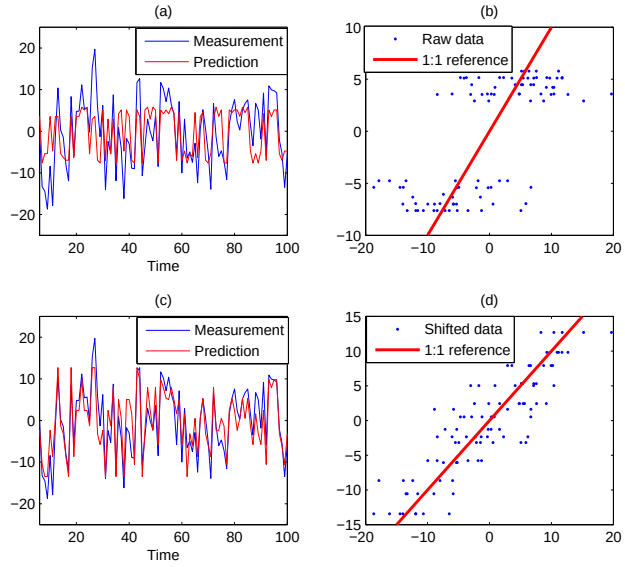


Figure 3.6: Comparison of PLS model performance for scenarios (1) ignoring time delays($R^2 = 0.44$): (a) and (b) ;(2) including time delays and shifting the inputs correspondingly($R^2 = 0.78$): (c) and (d)

Chapter 4

Integrated data cleaning and model identification

4.1 Motivation

As discussed in Chapter 2, although data cleaning methods can improve the overall data quality for further analysis,each process data set has unique features that make it hard to postulate a universally applicable data cleaning procedure. One challenge is how to determine whether the data have been cleaned enough after going through the cleaning step, i.e., the prevention of over-cleaning. In one approach certain improvements can be made on the data analytical tools such as soft sensors so that they are more intrinsically robust when dealing with low quality data [238]. For example, a robust version of partial least squares (RPLS)– the RSIMPLS algorithm [139] can handle contaminated data sets with outliers, and by incorporating the minimum covariance determinant estimator [256, 258], it can clean the data set and maximize the PLS model performance at the same time. Alternatively, inspired by the adaptive filter, we can combine the data cleaning step with a later stage of model building and performance evaluation to devise an integrated data cleaning scheme, as proposed in this chapter.

The remaining part of this chapter is organized as follows: preliminaries in the area of process model building and data cleaning are presented in Section 4.2. This is followed by the introduction of an adaptive filter and integrated data cleaning scheme in Section 4.3. In Section 4.4, the performance of the new data cleaning structure is illustrated with an industrial example and compared with the RSIMPLS algorithm, followed by concluding remarks in Section 4.5.

4.2 Preliminaries

In the process industries, numerous types of models can be used to describe process changes and predict the quality parameters. In this paper, we focus on partial least squares (PLS) that are key to multivariate process monitoring, and briefly introduce relevant information in the following subsections.

4.2.1 Partial least squares

Due to the complexity of modern petrochemical facilities, it is hard to derive first-principles models of the entire process; thus, the data-driven models, such as the soft-sensors, are often valuable tools than the model-based ones in process monitoring and fault detection [168, 195, 64]. An typical example of soft-sensors is the partial least squares(PLS) model, which is based on a dimensionality reduction technique maximizing the covariance between the predictor matrix $\mathbf{X}$ and the predicted matrix $\mathbf{Y}$ for each principal component spanning the reduced space [64, 112]. The standard mathematical formula for

PLS model is shown in Eq. (4.1):

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} = \sum_{k=1}^{a} \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F} = \sum_{k=1}^{a} \mathbf{u}_k \mathbf{q}_k^T + \mathbf{F} = \mathbf{X}B + B_0\end{aligned} \tag{4.1}$$

where $\mathbf{X}$ is an $n \times m$ matrix of predictors; $\mathbf{Y}$ is an $n \times p$ matrix of responses; $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n]^T$ and $\mathbf{U}$ are $n \times a$ score matrices; $\mathbf{P}$ and $\mathbf{Q}$ are $m \times a$ and $p \times a$ orthogonal loading matrices respectively; $\mathbf{E}$ and $\mathbf{F}$ are error matrices assumed to be independently and identically distributed; $B = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_p]_{m \times p}, B_0 = \left[\mathbf{b}_{0,1}^T, \mathbf{b}_{0,2}^T, ..., \mathbf{b}_{0,n}^T\right]_{n \times p}^T$ are the regression coefficients.

The SIMPLS algorithm [80] has been widely used for estimating PLS regression components. However, it was shown that the algorithm is very sensitive to outliers in a simulation study [139], and by incorporating the minimum covariance determinant (MCD) estimator [256, 258], they proposed a robust SIMPLS algorithm which outperformed the original one in resisting outliers. The new algorithm calculates the score distance, orthogonal distance and standard residual, defined in Eqs. (4.2) to (4.4):

$$SD_{i(k)}^2 = \mathbf{t}_i^T S_t^{-1} \mathbf{t}_i \tag{4.2}$$

$$OD_{i(k)} = \|\hat{\mathbf{x}}_i - \mathbf{P}\mathbf{t}_\mathbf{i}\| \tag{4.3}$$

$$resid = \frac{r_{i(k)}}{s_e} = \frac{y_i - b_{0,i} - \mathbf{x}_i B}{S_y - B^T S_x B} (if\ p = 1) \tag{4.4}$$

where k is the number of principal components included in the model; S_t, S_x, S_y are the covariance matrices of t-,x-,y-variables; $\hat{\mathbf{x}}_i$ is the robustly centered observations.

Observations will be considered as abnormal if the following conditions stand for each metric:

$$SD_{i(k)} > \sqrt{\chi^2_{k,0,975}} \tag{4.5}$$

$$OD_i > \sqrt{\hat{\mu}_{od^2} + \hat{\sigma}_{od^2} z_{0.975}} \tag{4.6}$$

$$|resid| > \sqrt{\chi^2_{1,0,975}} \tag{4.7}$$

where $\hat{\mu}_{od^2}$ and $\hat{\sigma}_{od^2}$ are the mean and standard deviation of the squared orthogonal distance; $z_{0.975} = \Phi^{-1}(0.975)$, the 97.5% quantile of Guassian distribution; $\chi^2_{k,\alpha}$ stands for the chi-square distribution.

In this section, we focus on tuning the moving window size (=2h+1) of the 3σ rule, the Hamel identifier and the Savitzky-Golay filter, the expressions of which are described in Eqs. (2.9), (2.10), (2.20) and (2.21).

4.3 Method description

4.3.1 Adaptive filter

Fig. 4.1 shows a block diagram of the adaptive filter where a sample from input signal $x(t)$ is fed into the filter block that calculates a corresponding output signal $y(t)$ at the same time point. The filter block contains parameters that can be tuned to ensure the performance of the filter. Generally speaking, the adaptive filter extracts a reference signal $d(t)$ from the input signal $x(t)$ by comparing it with the filtered signal $y(t)$ and feeds the difference $e(t)$ into the parameter updating block to tune the filter. The updating algorithms include

the Wiener solution, steepest descent, least mean squares (LMS), etc. [127]. Generally, an adaptive filter contains four important components [88]:

- Signals to be processed
- A structure containing methods to calculate output signals of the filter from the input signals
- Parameters within the structure that can be tuned to change the input-output relationship of the filter
- An adaptive algorithm that can be used to update the parameters

Applications of adaptive filters include system identification [19], inverse modeling [319], linear prediction [196] and feed forward control [88].

4.3.2 Integrated data cleaning scheme

Based on the adaptive filter framework, an integrated data cleaning scheme is proposed as shown in Fig.4.2. The diagram consists of six components: raw data, clean test data, final clean data, cleaning methods, model building and parameter update algorithm. Similar to the adaptive filter, the data cleaning scheme is defined by six important attributes:

- Raw data to be processed
- Clean test data to be used to validate model performance
- A structure containing methods to clean the raw data

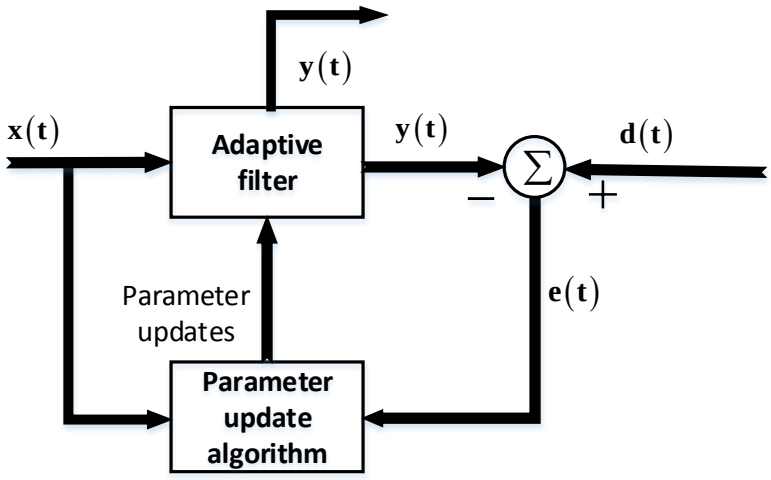


Figure 4.1: Diagram of adaptive filtering

- A model building and validation step
- Parameters within the structure that can be tuned to control the aggressiveness of the data cleaning methods
- An adaptive algorithm that can be used to update the parameters

A mathematical framework for above scheme is shown in Eqs. (4.8)-(4.12):

$$\hat{\mathbf{X}} = f\left(\mathbf{X}, \theta_{k+1}\right) \tag{4.8}$$

$$q = h\left(\hat{\mathbf{X}}, \mathbf{Z}\right) \tag{4.9}$$

$$e = \bar{q} - q \tag{4.10}$$

$$\theta_{k+1} = g\left(\theta_k, e\right) \tag{4.11}$$

$$\mathbf{Y} = f\left(\theta_{final}, \mathbf{X}\right) \tag{4.12}$$

where $\mathbf{X}$ stands for the raw data, $\hat{\mathbf{X}}$ and $\mathbf{Y}$ stand for the temporary and final cleaned data sets respectively. θ_k and θ_{final} represent the intermediate (at kth iteration) and final parameter settings for the data cleaning methods respectively. q represents the model quality statistics and $\bar{q}$ is the threshold set for q. f, h, g are function sets for data cleaning, model building and validation and parameter updating respectively.

After going through a preliminary cleaning step, the data $\hat{\mathbf{X}}$ is fed into the model building block that constructs dynamic models, partial least squares (PLS) models, etc. The performance of the model is tested using a clean test data set $\mathbf{Z}$ either coming from a normal operation batch (the "golden"batch) or a physical model simulation. The metric q evaluating model performance (R^2 for example) is compared with a reference $\bar{q}$, and the resulting difference e is sent to the parameter update algorithm block to tune parameters θ_k of the data cleaning methods, where the grid search is performed. After a series of iterations, we can obtain a final clean data set $\mathbf{Y}$ and satisfactory model performance at the same time.

4.4 Case study

The applicability of the proposed data cleaning scheme is illustrated with data from an industrial case. The data set was used for studying outlier detection and noise removal, and a PLS model was built in the model performance block.

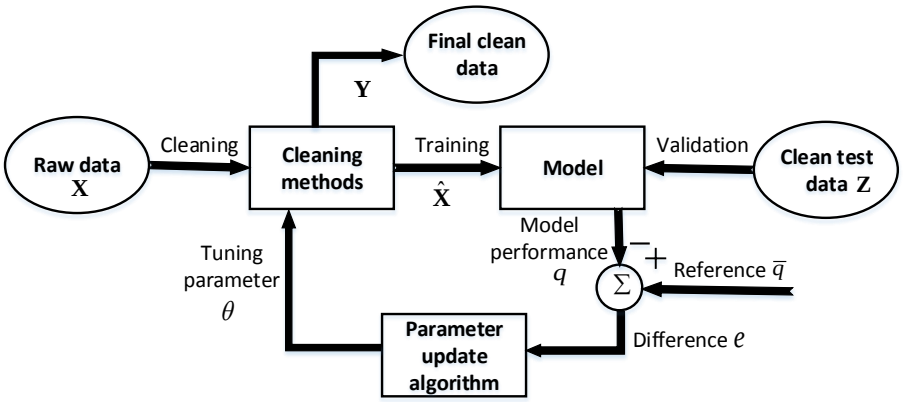


Figure 4.2: Diagram of adaptive cleaning procedure

4.4.1 Dataset

The slurry-fed ceramic melter (SFCM) is used for solidifying the liquid wastes produced during the reprocessing of nuclear fuels, and a simple schematic of the SFCM operated at West Valley is shown in Fig. 4.3. The slurry composed of glass formers and waste is spread onto the surface of the molten glass pool, and a few electrodes are used for heating the mixture. The dried feed left forms a "crust"or "cold cap"melting continuously into the glass phase. Glass is discharged periodically from the melter leading to periodic fluctuations of the glass level. As shown in Fig. 4.3, the temperatures inside the melter are monitored extensively at 20 locations and used as process inputs. Moreover, the glass tank level data are also collected as the process output. The melter data are sampled every 5 minutes because of a long process response time to step changes usually on the order of hours [324].

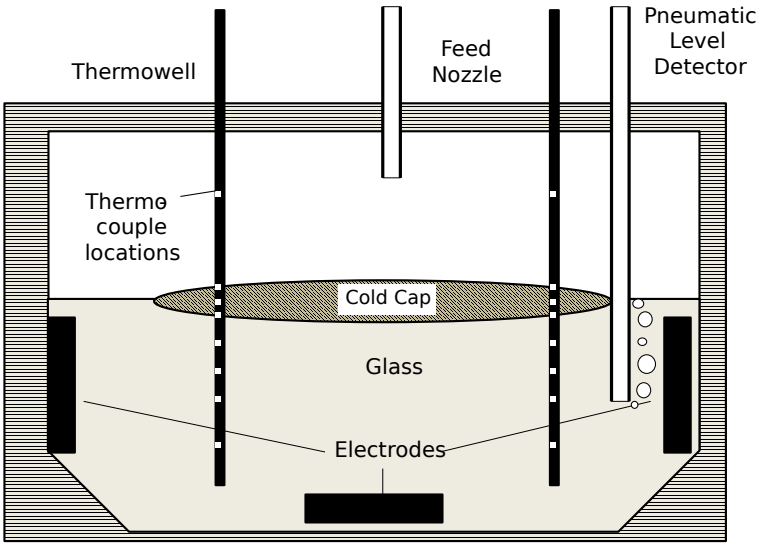


Figure 4.3: Schematic drawing of slurry-fed ceramic melter [324]

The clean training data set and testing data set are shown in Figs. 4.4 and 4.5. Random noise following a standard normal distribution was added to the glass level to degrade the quality of the training data set, and the resulting noisy data set is shown in Fig. 4.6. The outlier contaminated data set was simulated by adding one outlier per every 20 samples starting from 20th observation with an equal probability of being either $+1.5$ or -1.5, as shown in Fig. 4.7. Both Figs. 4.6 and 4.7 exhibit a wider and more irregular dispersion in level axis than Fig. 4.4.

4.4.2 Outlier detection

- **RSIMPLS algorithm**

The RSIMPLS algorithms was implemented on the outlier contaminated

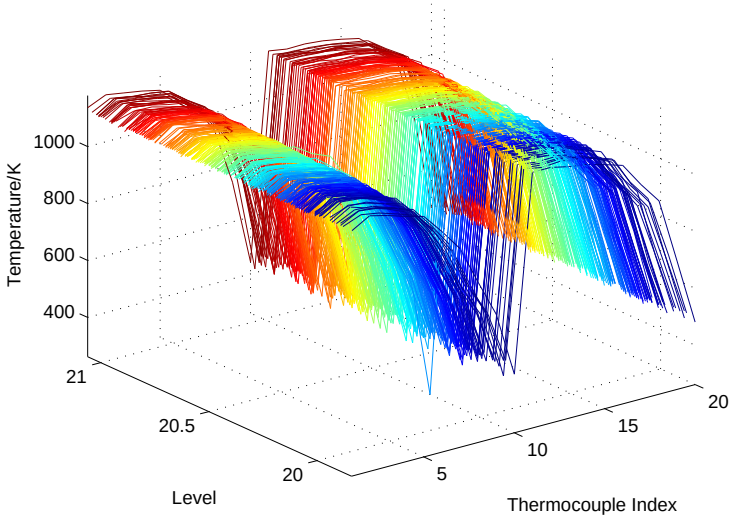


Figure 4.4: SFCM clean training data set

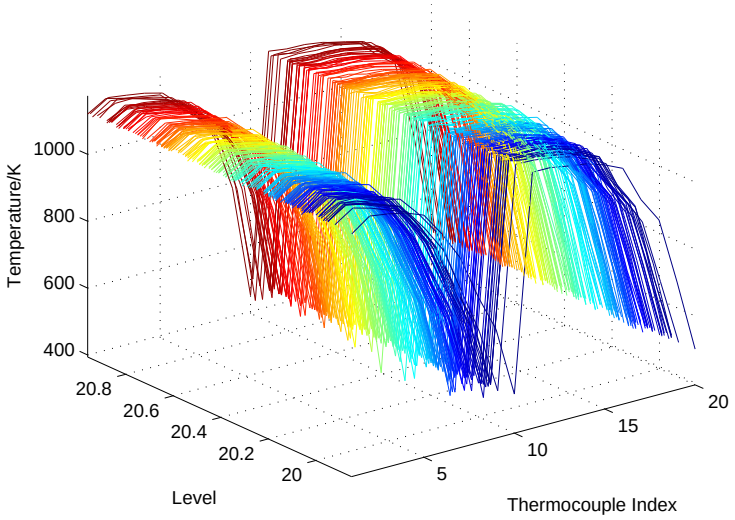


Figure 4.5: SFCM clean test data set

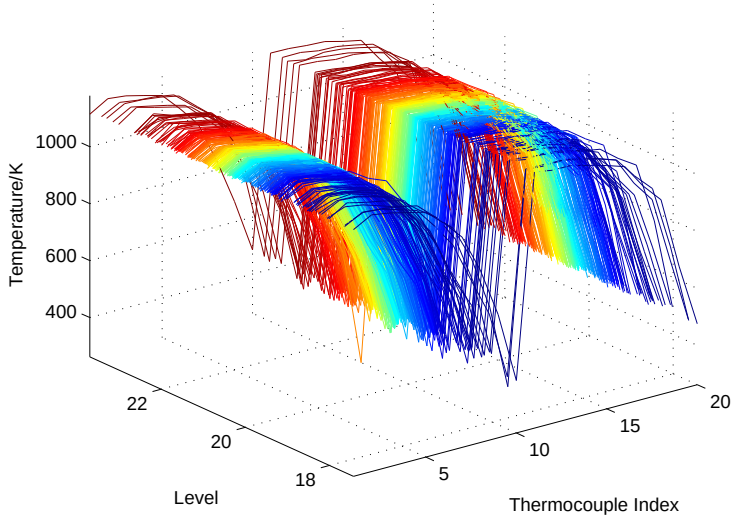


Figure 4.6: SFCM noisy training data set

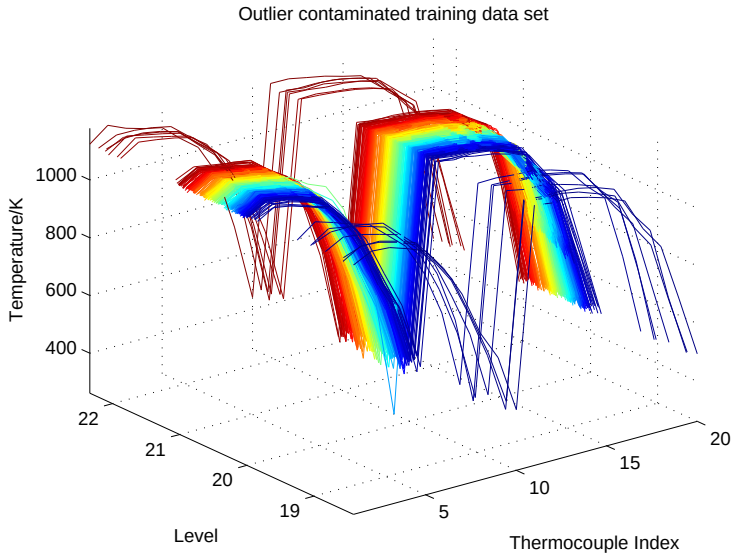


Figure 4.7: SFCM outlier contaminated training data set

training data set, and a PLS model with three principal components was selected (k=3). Figs. 4.8 and 4.9 show the score and regression diagnostic plots using RSIMPLS algorithm. The red lines represent the thresholds calculated from Eqs. 4.5 to 4.7, and indices of most deviated points are marked in the figure. Combining Figs.4.8 with 4.9, we can see that the RSIMPLS algorithm only picks three true outliers with indices of 20, 100 and 160. A possible explanation is that the RSIMPLS algorithm detects outliers by exploring the correlation between predictor matrix $\mathbf{X}$ and the predicted matrix $\mathbf{Y}$, and the changes in $\mathbf{X}$ may mask the existence of outliers in $\mathbf{Y}$.

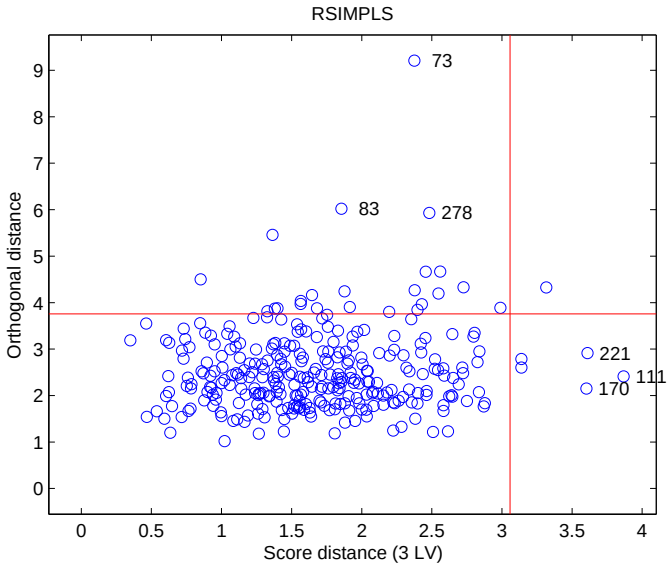


Figure 4.8: Score diagnostic plots using RSIMPLS algorithm

- **The Integrated data cleaning scheme**

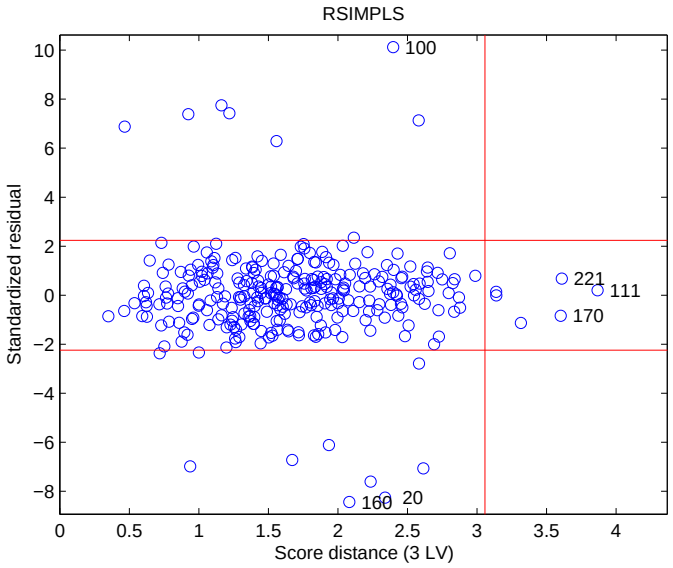


Figure 4.9: Regression diagnostic plots using RSIMPLS algorithm

The training data containing outliers were fed into the adaptive data cleaning module shown in Fig. 4.2 and after setting the reference $R^2 = 0.82$, we obtained a better outlier detection results shown in Figs. 4.10 and 4.11: the 3σ rule missed outliers with indices of 40, 160 and 220, and the Hampel identifier missed point 160 and 280; however, they were able to detect most of the added outliers.

Figs. 4.12 and 4.13 illustrate changes of the PLS model performance R^2 with an increasing moving window size during the iterative updating process: the wavy curves show that the models achieve the best performance at $h = 8$ for the 3σ rule and $h = 14$ for the Hampel identifier, corresponding to window sizes of 17 and 29 respectively.

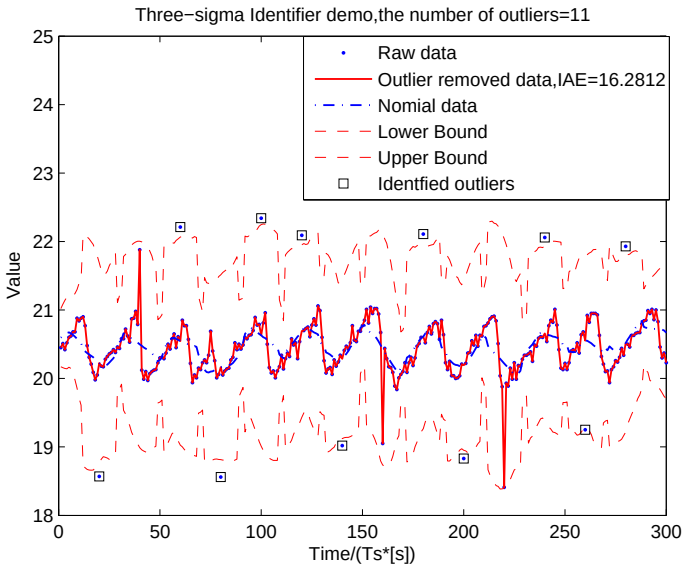


Figure 4.10: Outlier removed SFCM level data, 3σ rule, window size=17

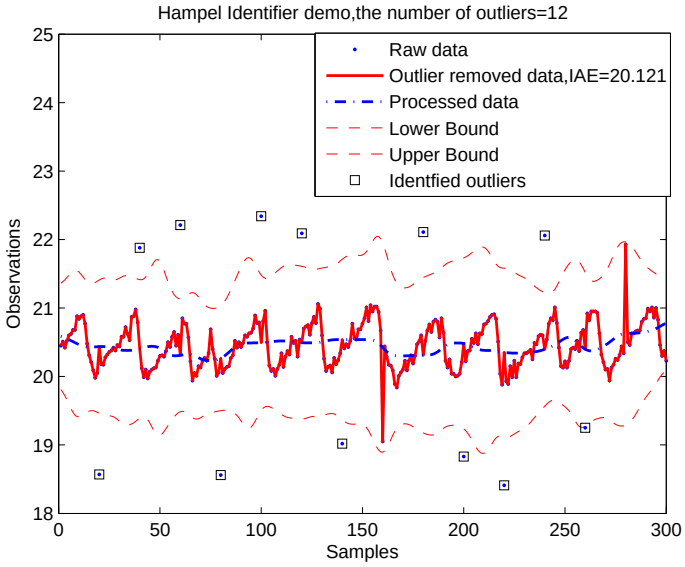


Figure 4.11: Outlier removed SFCM level data, Hampel identifier, window size=29

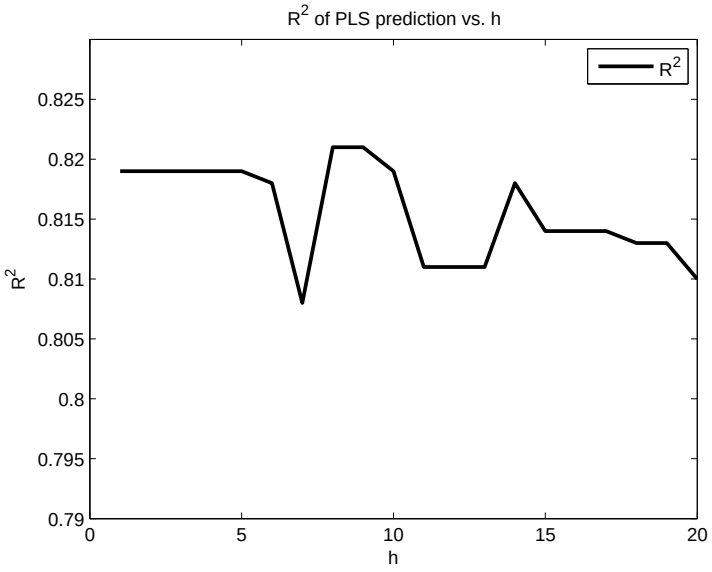


Figure 4.12: PLS model performance R^2 changes with a increasing moving window size of 3σ rule

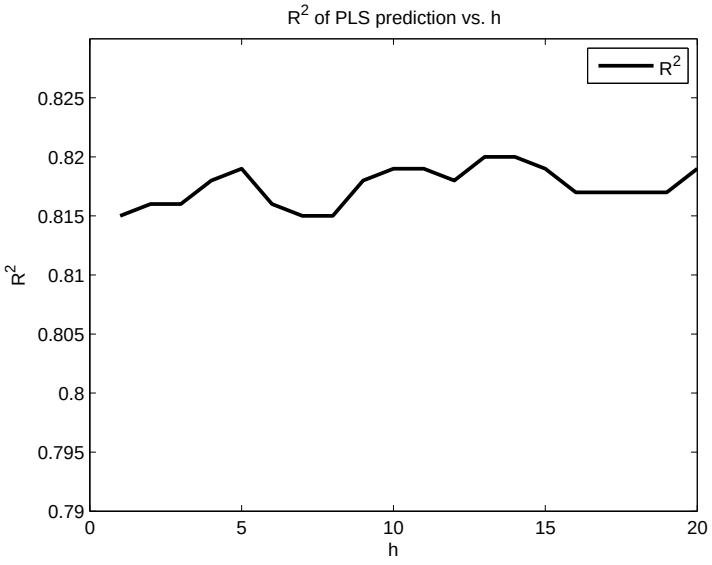


Figure 4.13: PLS model performance R^2 changes with a increasing moving window size of Hampel identifier

A summary of the PLS model test results is shown in Table 4.1:

Table 4.1: PLS test results

RSIMPLS	Integrated data cleaning scheme	
	3σ rule	Hampel identifier
0.81	0.82	0.82

As shown in above table, the RSIMPLS algorithm and integrated data cleaning scheme obtain close results, probably because of a very limited influence of outliers in this case. However, considering the successful outlier detection rate, the data cleaning scheme seems to be a better choice.

4.4.3 Noise removal

A PLS model was built using noisy training data set and the prediction result is shown in Fig. 4.14. After injecting the noisy training data set into the adaptive data cleaning module shown in Fig. 4.2 and setting the reference $R^2 = 0.68$, we obtained a filtered data set with an ensured model performance shown in Fig. 4.15.

Fig. 4.16 illustrates changes of the PLS model performance R^2 with an increasing moving window size during the iterative updating process: the parabola-shaped curve shows that the model achieves the best performance at h=2 (i.e, window size=5) and degrades significantly after h=8, implying that an over-filtering problem occurs.

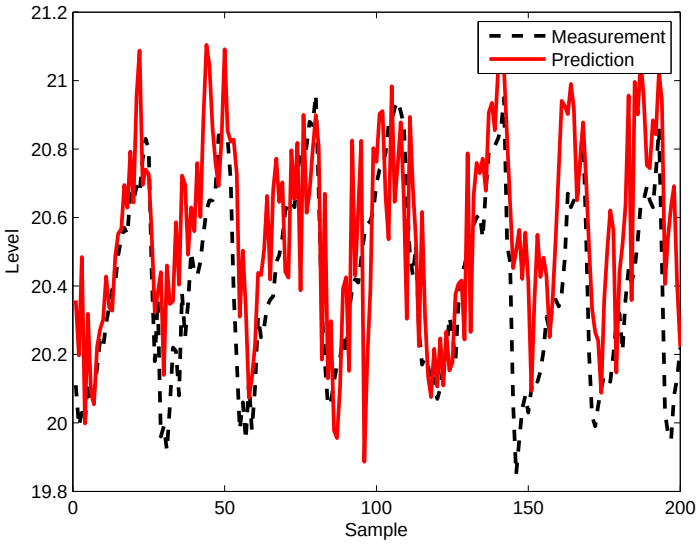


Figure 4.14: Level prediction based on noisy SFCM data, $R^2 = 0.46$

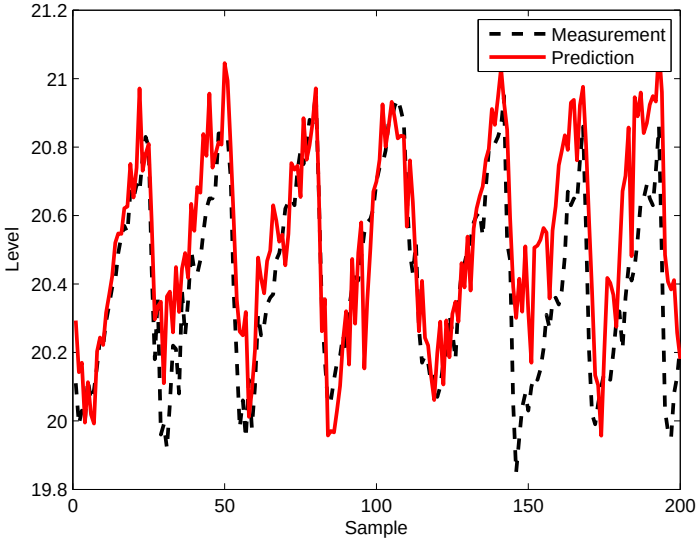


Figure 4.15: Level prediction based on filtered SFCM data, window size=5, $R^2 = 0.69$

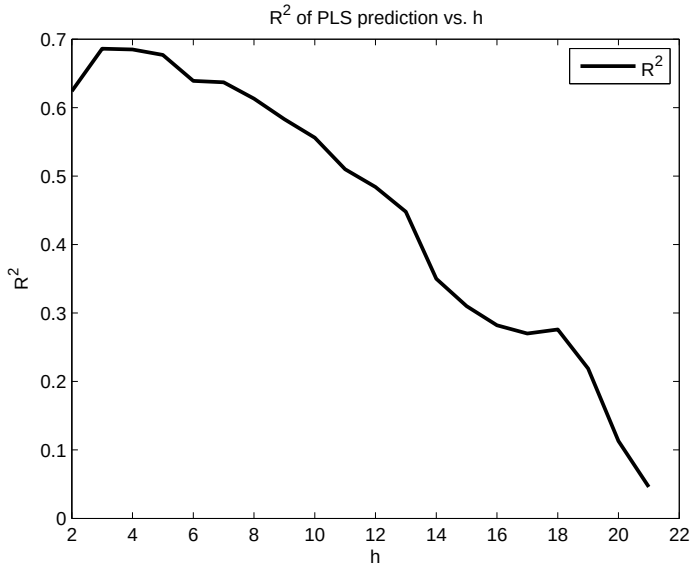


Figure 4.16: PLS model performance R^2 changes with a increasing moving window size of SG filter

4.5 Summary

In this chapter, an integrated data cleaning scheme is proposed, and it includes data cleaning, parameter tuning and model performance estimation. In comparison with other methods treating a low-quality data set, the scheme enjoys more versatility: for a given model building task, we can substitute different data cleaning techniques and the scheme can automatically tune the parameters based on the performance of the model. A case study based on a industrial process was conducted. The results show that the integrated data cleaning scheme can circumvent the over-cleaning problem by incorporating the model performance evaluation, and simultaneously obtain a clean data set and a satisfactory model.

In this chapter, we have only applied the univariate data cleaning methods such as the 3σ rule and the Hampel identifier. Moreover, the parameter updating algorithm used to find the optimal model performance is not very effective. Research in progress is focused on applying multivariate data cleaning methods and a more efficient optimality searching algorithm.

Chapter 5

An improved methodology for outlier detection in dynamic data sets

5.1 Motivation

As mentioned in Chapter 2, a key assumption underlying many statistical outlier detection methods[120, 26, 255, 256, 231]is that the data are identically and independently distributed (*i.i.d.*), which is often compromised by dynamics hidden in time-varying data sets [1]. A related issue is how to differentiate process dynamics and outliers. Traditionally, a moving window technique is applied, which assumes that data in a small enough moving window can still be treated as identically and independently distributed. However, applying moving window techniques does not always give satisfactory results, especially when the variations in the data set are significant, and it is always computationally expensive for large data sets. Another solution is approximating variations in the data by time-series models, such as an autoregressive model (AR), and then separating observations that are inconsistent with the imposed model using outlier detection techniques proposed in the time series literature,

[1] Xu S, Baldea M, Edgar TF, Wojsznis W, Blevins T, Nixon M. An improved methodology for outlier detection in dynamic datasets. AIChE Journal. 2015;61(2):419-433.The project was supervised by Dr. Michael Baldea and Dr. Thomas F. Edgar. Willy Wojsznis, Terry Blevins and Mark Nixon gave technical support and conceptual advice

such as likelihood-based methods[103, 54, 297, 56, 104, 298, 148, 192]. However, those methods are only applicable to small data sets with a small number of outliers; besides, there is a lack of discussion on industrial applications and related parameter tuning issues. The on-line filter-cleaner [186] has been applied to univariate outlier detection in an industry data set, and outperforms other time series outlier detection methods in robustness and efficiency. However, the filter-cleaner[186] can still be improved in several aspects such as the model fitting algorithm, parameter tuning, and an extension to the multivariate cases. Besides, since off-line data analyses are more encountered in practice, it is useful to develop a related off-line outlier detection method. Finally, detecting innovational outliers, which are commonly encountered in dynamic processes, is also worth studying .

5.2 Preliminaries

5.2.1 Contamination rate,detection rate,mis-identification rate,and normal data estimation rate

The contamination rate κ is defined as the percentage of outliers in total data, and the detection rate χ is defined as the percentage of outliers being successfully identified. The mis-identification rate β is defined as the percentage of normal data falsely tagged as outliers (type I error), and γ is defined as a prior estimation of the percentage of normal data in the original data set. Normally, γ is set to be larger than 80 %; otherwise, the data will be considered as of poor quality and useless. Mathematical expressions for κ,

χ, β and γ are shown in Eqs. (5.1) to (5.4).

$$\kappa = \frac{N_{outliers}}{N_{total\ data}} \tag{5.1}$$

$$\chi = \frac{N_{successfully\ identified}}{N_{total\ outliers}} \tag{5.2}$$

$$\beta = \frac{N_{false\ alarm}}{N_{normal\ data}} \tag{5.3}$$

$$\gamma = \frac{N_{normal\ data}}{N_{total\ data}} \tag{5.4}$$

Based on the definition, $\kappa = 5\%$ in a data set with 10000 observations means the number of actual outliers is 500. To simplify the verification process of detection methods, we are using simulated data sets with outliers added at every 20th sample points.

5.2.2 Principal component analysis(PCA) and dynamic PCA

The principal component analysis (PCA) method has been applied to detect multivariate outliers. First, PCA decomposes the original data set using singular value decomposition shown in Eq. (5.5). Second, it calculates the Hotelling's T^2 metric defined in Eq. (5.6), and by monitoring such a metric, disturbances or abnormalities can be detected and isolated if the metrics violate the threshold calculated in Eq. (5.7).

$$\mathbf{X_{PCA}} = \mathbf{U\Sigma V^T} \tag{5.5}$$

where $\mathbf{X}$ is the training data matrix with n observations and m variables.

$$t^2 = \mathbf{x}^T V \Sigma_a^{-2} V^T \mathbf{x} \tag{5.6}$$

where V contains the loading vectors associated with the a largest singular values, Σ_a includes the first a rows and columns of Σ, and $\mathbf{x}$ is an observation vector of dimension m.

$$T_\alpha^2 = \frac{(n-1)\,a}{(n-a)} F_\alpha\,(a, n-a) \tag{5.7}$$

where $F_\alpha\,(a, n-a)$ is the critical point of the F-distribution with a and $n-a$ degrees of freedom, and α is the level of significance.

However, the standard PCA method fails to take into consideration the serial correlation at different time instances, and the dynamic principal component analysis (DPCA)[172, 260] has been proposed for detection and isolation of process disturbances in time series data. The DPCA augments each observation $\mathbf{X}$ matrix at time t with the previous l time instances, as shown in Eq. (5.8), and similar to PCA, it decomposes $\mathbf{X_{DPCA}}$ and detects outliers by monitoring the Hotelling's T^2.

$$\mathbf{X}_{DPCA}\,(l) = [\mathbf{X}\,(t)\,\mathbf{X}\,(t-1)\cdots\mathbf{X}\,(t-l)] \tag{5.8}$$

where $\mathbf{X}\,(\mathbf{t})$ is the observation matrix at time instance t.

5.3 Time series Kalman filter

5.3.1 Univariate autoregressive (AR) model fitting

In practice, many discrete random processes can be approximated by a stationary ARMA (p,q) model shown in Eq. (5.9)

$$\left(1 + \sum_{i=1}^{p} \phi_i z^{-i}\right) x_t = \left(1 + \sum_{i=1}^{q} \theta_i z^{-i}\right) \epsilon_t \tag{5.9}$$

where p,q are model orders, ϕ_i, θ_i are coefficients, ϵ_i is the white noise of the model, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and z^{-1} is a shift operator[43].

Based on the Wold decomposition theorem [325] and Kolmogorovs theorem[166], any ARMA process can be represented by an AR process of infinite order. Thus, a feasible solution is to fit an AR (p) model as shown in Eq. (5.10) in order to approximate the changes exhibited in the ARMA process.

$$\left(1 + \sum_{i=1}^{p} \phi_i z^{-i}\right) x_t = \epsilon_t \tag{5.10}$$

Three different methods are commonly used in estimating ϕ_i in AR(p) model shown in Eq. (5.10): the least-squares method, the Yule-Walker method, and Burg's method[159]. The previous two methods involve an inverse of the auto-covariance matrix step, while the Burg's method calculates the reflection coefficients and then applies Levinson recursion to obtain the AR parameter estimates. De Hoon et al.[79] compared the above methods and presented simulation results showing that because the first two methods invert an auto-covariance matrices which can be poorly conditioned, Burg's method is preferable to the least-squares and the Yule-Walker approach.

5.3.2 Multivariate (vector) autoregressive (MVAR) model fitting

For MVAR model fitting, an extension of Yule-Walker method to multivariate cases can be applied[200], as well as the multivariate Burg's method[223, 224]. A new estimator (Arfit) has been proposed[220, 267], and by comparing the above multivariate estimators, the multivariate Burg's method still

outperforms the others[201, 266].

5.3.3 Model order selection

When selecting a model order, a balance has to be made in between improving the coefficient of determination R^2 and a prevention of model overfitting. For AR model order selection, a commonly used criterion is the Akaike information criterion (AIC)[6] shown in Eq. (5.11):

$$AIC\left(p\right) = N \log \hat{\rho}_p + 2p \tag{5.11}$$

Another way to select model order is the Schwarz's Bayesian information criterion (BIC) shown in Eq. (5.12) [270], which can be applied in both AR and MVAR models.

$$BIC\left(p\right) = \frac{l_p}{m} - \left(1 - \frac{n_p}{N}\right) \log N \tag{5.12}$$

where p is the model order, m is the number of variables, N is the number of observations, n_p is the number of model parameters, and

$$l_p = \log \left\{ \det \left[\left(N - n_p\right) \hat{\boldsymbol{\rho}}_p \right] \right\} \tag{5.13}$$

$\hat{\boldsymbol{\rho}}_p$ stands for residual covariance matrix, $\det\left[.\right]$ calculates the matrix's determinant. The QR factorization algorithm is applied in evaluating $\hat{\boldsymbol{\rho}}_p$, and a regularization term δD^2, where δ is a coefficient and D^2 is a positive definite diagonal matrix, is added to deal with the situation when $\hat{\boldsymbol{\rho}}_p$ becomes ill-conditioned[220, 267].

For the univariate case, m=1, and

$$l_p = \log \{|(N - n_p)\, \hat{\rho}_p|\} \tag{5.14}$$

Because no AIC calculation equation has been found for multivariate cases, the BIC is used in the new algorithm. Normally, the best model order corresponds to the lowest BIC or AIC value. The outliers will contaminate the data set and give rise to a large model order p. Thus, a pre-whitening procedure is needed to reduce the outlier effects on model order estimation. After pre-whitening, a small p usually suffices, similar results have been reported[186]. In this dissertation's simulation study, p is selected to be 2.

5.3.4 Combining time series modeling with Kalman filter

Inspired by the filter-cleaner[186](shown in Appendix B), the time series Kalman filter(TSKF) is proposed which differs in several aspects:

1. Burg's model estimation is applied. In multivariate cases, the auto-covariance matrices might become ill-conditioned; thus, Burg's method is preferred to the Yule-Walker method.

2. Parameter estimation is directly obtained from the preliminary clean data set, which reduces efforts to get a robust estimation of the auto-covariance matrix via reweighted MCD.

3. Instead of applying the filter-cleaner[205] which simultaneously detects and replaces outliers with related predicted values, the detection and

clean steps can be separated: for on-line use, the users can do both at the same time, and for off-line use, they also can replace the outliers after the detection process is finished, options are available for applying the best imputation techniques to replace the outliers for a specific application. In the new procedure, for simplicity, neighboring normal points will be used to replace outliers instead of using a model predicted value, because in practice, no actual model exists; imposing a fitted model, though a robust one, will compromise the integrity of the original data set and might lead to a new spike such as shown in Fig. 5.1 (close to time point 600).

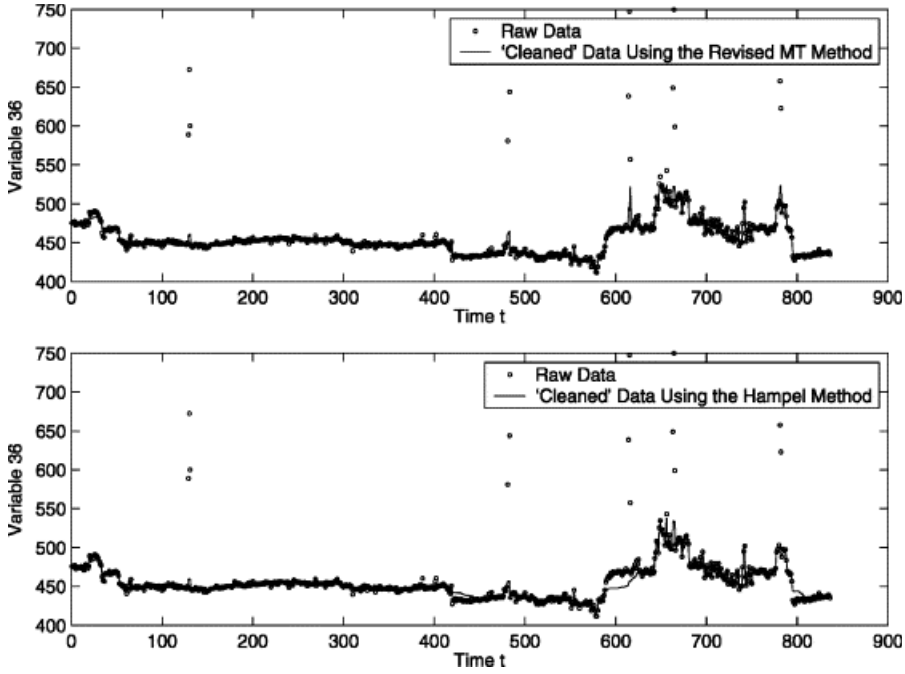


Figure 5.1: Testing results of filter-cleaner and the Hampel method[186]

4. Besides the additive outliers, the simulation cases include innovational outliers.

Both off-line and on-line versions of the method are provided. For simplicity, the method is written in the format for multivariate outlier detection.

5.3.4.1 Off-line version

Given a data set $\{\mathbf{y}_t\}_{t=1}^{N}$, we can detect and replace the outliers with the following steps:

1. **Data partition**: partition the data set into M subsets,$\{\mathbf{y}_t\}_{t=1}^{N_i}, i = 1, 2, ..., M$.

2. **Pre-whitening**: for each subset$\{\mathbf{y}_t\}_{t=1}^{N_i}, i = 1, 2, ..., M$, pre-whiten the data using reweighted MCD estimator, replace the outliers with robust center $\boldsymbol{\mu}_i$, and centralize the data with $\boldsymbol{\mu}_i$.

3. **Model fitting**: based on the preliminary clean data $\{\mathbf{y}_t^c\}_{t=1}^{N_i}$,

3.1. (Optional) select the model order p according to BIC.

3.2. Calculate the model coefficients $\boldsymbol{\Phi}_i$ based on Burg's method.

4. **Outlier detection**: for each subset$\{\mathbf{y}_t\}_{t=1}^{N_i}, i = 1, 2, ..., M$:

4.1. Reformat:

$$\begin{aligned} \mathbf{Y}_t &= \Theta \mathbf{Y}_{t-1} + \mathbf{U}_t \\ \mathbf{y}_t &= \mathbf{H}\mathbf{Y}_t \end{aligned} \tag{5.15}$$

where

$$\mathbf{Y}_t^T = \left[\mathbf{y}_t, \mathbf{y}_{t-1}, ..., \mathbf{y}_{t-p+1}\right]_{1\times pm} \tag{5.16}$$

$$\mathbf{U}_t^T = \left[\hat{\boldsymbol{\epsilon}}, \mathbf{0}, ..., \mathbf{0}\right]_{1\times pm}; \hat{\boldsymbol{\epsilon}} \sim N\left(\mathbf{0}, \mathbf{Q}\right) \tag{5.17}$$

$$\mathbf{H} = [\mathbf{I}_{m\times m}, \mathbf{0}, ..., \mathbf{0}]_{1\times pm} \tag{5.18}$$

$$\Theta = \begin{bmatrix} \mathbf{\Phi}_{1,m\times m} & \mathbf{\Phi}_{2,m\times m} & \cdots & \mathbf{\Phi}_{p-1,m\times m} & \mathbf{\Phi}_{p,m\times m} \\ \mathbf{I}_{m\times m} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m\times m} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{m\times m} & \mathbf{0} \end{bmatrix}_{pm\times pm} \tag{5.19}$$

4.2 Predict:

$$\hat{\mathbf{Y}}_{t|t-1} = \Theta\hat{\mathbf{Y}}_{t-1|t-1} \tag{5.20}$$

$$\mathbf{P}_{t|t-1} = \Theta\mathbf{P}_{t-1|t-1}\Theta^T + \mathbf{Q} \tag{5.21}$$

4.3 Update:

$$\mathbf{E}_t = \mathbf{y}_t - \mathbf{H}\hat{\mathbf{Y}}_{t|t-1} \tag{5.22}$$

$$\mathbf{S}_t = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^T + \tau\mathbf{I} \tag{5.23}$$

$$d_t = \sqrt{\mathbf{E}_t^T\mathbf{S}_t^{-1}\mathbf{E}_t} \tag{5.24}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^T\mathbf{S}_t^{-1} \tag{5.25}$$

$$\hat{\mathbf{Y}}_{t|t} = \hat{\mathbf{Y}}_{t|t-1} + \mathbf{K}_t\mathbf{E}_t \tag{5.26}$$

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K_t}\mathbf{H})\,\mathbf{P}_{t|t-1} \tag{5.27}$$

4.4 Detect:

4.4.1. Set $\Delta = 0$.

4.4.2. Find a number of n observations whose Mahalanobis distance

$d_t \geq \Delta$.

4.4.3. Calculate the percentage of normal data

$$\xi = \frac{N^i - n}{N^i} \tag{5.28}$$

4.4.4. If $\xi \geq \gamma$,stop; else increase Δ by $d\Delta$.

4.4.5. The outliers correspond to observations with Mahalanobis distance

$$d_t \geq \Delta_{final}.$$

4.5 Replace: replace the outliers with neighboring normal values.

5.3.4.2 On-line version

The on-line version of the method is very similar to the on-line filter-cleaner[186]:

Given a data sequence at time , we can detect and replace the outliers in the following steps:

1. Choose a data set $\{\mathbf{y}_t\}_{t-M+1:t}^{M}$ with window size MW.

2. The pre-whitening and modeling fitting steps are the same as the off-line version.

3. Based on a pre-set threshold Δ, if the Mahalanobis distance $d_t \geq \Delta$, the observation is identified as an outlier.

4. Replace the outliers with neighboring normal values.

The new procedure keeps most of the original structure of the Kalman filter[155] unchanged, and only makes a few modifications shown in Eqs. (5.23)

to (5.24). In Eq. (5.23), the variance of observation noise term $\mathbf{R}$ shown in Eq. (5.30) of the original Kalman filter is deleted, which makes the new algorithm detect outliers without changing the original observations. A Tikhonov regularization term [293] $\tau\mathbf{I}$ is added to deal with ill-conditioned covariance matrices in multivariate outlier detection cases. If we add the observation noise terms back, Eqs. (5.15) and (5.23) become:

$$\begin{aligned} \mathbf{Y}_t &= \Theta\mathbf{Y}_{t-1} + \mathbf{U}_t \\ \mathbf{y}_t &= \mathbf{H}\mathbf{Y}_t + \mathbf{z}_t \end{aligned} \tag{5.29}$$

$$\mathbf{S}_t = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^T + \mathbf{R} \tag{5.30}$$

where $\mathbf{z}_t \sim N(\mathbf{0}, \mathbf{R})$.

A univariate version of the method is similar, except that a univariate Mahalanobis distance is calculated as shown in Eq. (5.31):

$$d_t = |e_t| / \sqrt{s_t} \tag{5.31}$$

As we can see from the method description shown above, the TSKF method has a high computational cost largely due to the outlier detection step, which tracks the variances of each observation point. Such a step becomes even slower when dealing with multivariate outlier detection cases.

5.3.4.3 Parameter tuning

An important evaluation standard for an outlier detection method is whether contains tuning parameters. Table 5.1 gives the tuning parameters for the implementation of the TSKF method.

Table 5.1: Tuning parameters of the TSKF method

Implementation	Tuning parameters
on-line	distance threshold Δ; moving window size(MW)
off-line	γ; partition window size(PW)

For on-line use, since we have no knowledge of how to pre-set the distance threshold Δ, we need to run a training data set and record the Mahalanobis distance d_t, which can be used to set the value of Δ.

For off-line use, we can set the γ based on prior knowledge of the raw data set, such as $\gamma = 95\%$ means we estimate that the maximum amount of outliers in the data set should not exceed 5 % of the total number of observations. γ can be tuned easily without repeating the recursion calculation of the Mahalanobis distance d_t.

5.4 Simulation Testing

For illustration, the on-line and off-line versions of the TSKF method will be applied to both univariate and multivariate outlier detection in the dynamic data sets. For simplicity, only the univariate simulation process is described: following the approach used by Liu et al.[186], we obtain data by simulating the additive outlier (AO) and innovational outlier(IO) models shown in Eq. (5.32), and Eq. (5.33); v_t has equal probabilities being $+Amp$ or $-Amp$ (Amp is the amplitude of outliers); x_t follows ARMA(1,1) processes for univariate cases. The multivariate simulation process is the similar as the

univariate one.

$$y_t = v_t \delta_t^{(T)} + x_t \tag{5.32}$$

$$y_t = \frac{\left(1 + \sum_{i=1}^{q} \theta_i z^{-i}\right)}{\left(1 + \sum_{i=1}^{p} \phi_i z^{-i}\right)} v_t \delta_t^{(T)} + x_t \tag{5.33}$$

where $\delta_t^{(T)}$ is a pulse function: $\delta_t^{(T)} = 1$ if $t = T$; $\delta_t^{(T)} = 0$ if $t \neq T$. v_t is an outlier. From the above equations we can see that additive outliers (AOs) affect the observed time series only at time T, while the innovational outliers (IOs) impact a finite number of observations in a stationary process.

The univariate model was run with 10000 test points for both on-line and off-line cases. For processes with additive outliers(AOs), the data are corrupted with outliers at different contamination rates defined in Eq. (5.1). For processes with innovational outliers (IOs) cases, IOs are added every 100th sample points in ARMA(1,1) processes.

The multivariate models are run with 10000 test points for off-line case and with 1000 test points for on-line case. Similar to the univariate model, data are contaminated with different rates of outliers for the AO process simulation. For processes with innovational outliers (IOs) cases, IOs are added every 100th sample points in VARMA(1,1) processes.

A summation of simulation cases is shown in Table 5.2:

In addition, for on-line testing, the moving window size(MW)is set to

Table 5.2: Brief summation of simulation cases

Models	AO	IO	Test points number on-line	off-line
ARMA(1,1)	✓	✓	10000	10000
VARMA(1,1)	✓	✓	1000	10000

be 100, and for off-line testing, the partition window size (PW) is set to be 1000.

Furthermore, although the definitions of outlier detection rate χ and mis-identification rate β work perfectly for additive outlier (AO) detection, some modifications need to be made, as shown in Eqs. (5.34) and (5.35), so that they can be applied to innovational outlier(IO) detection.

$$\chi_{IO} = \frac{N_{successfully\ identified\ \delta_t^{(T)}=1}}{N_{total\ \delta_t^{(T)}=1}} \tag{5.34}$$

$$\beta_{IO} = \frac{N_{false\ alarm\ prior\ to\ \delta_t^{(T)}=1}}{N_{normal\ data}} \tag{5.35}$$

As shown in Eq. (5.34), the χ_{IO} is defined as the percentage of successfully identified locations of pulse function ($\delta_t^{(T)} = 1$) leading to the innovational outliers. In simulation cases, the original data set is divided into subsets with 100 observations each, and one pulse function is added on each subset. Based on the definition of β_{IO} shown in Eq. (5.35), any identified outliers prior to the pre-set locations of pulse function $\delta_t^{(T)} = 1$ within each subset will be regarded as false alarms.

5.4.1 Model impact analysis and order selection

The cleanness of the preliminary clean data will affect the performance of the outlier detection, and to demonstrate the necessity of a pre-whitening step, a detailed discussion is given based on the simulation case: an ARMA(1,1) model with $\phi = 0.9, \theta = 0$, and $Amp = 4$.

The true model of the process is expressed as Eq. (5.36):

$$\left(1 - 0.9z^{-1}\right) x_t = \epsilon_t \tag{5.36}$$

We first estimate the model order based on BIC shown in Eq. (5.12) for raw data of single moving window size(on-line) and pre-whitened data of single partition window size(off-line). Fig. 5.2 shows that the BIC results are not significantly affected by the pre-whitening step; also a lower model order of 1 or 2 usually suffices. Thus, to prevent model over-fitting, the order is selected to be 2 for both on-line and off-line implementation.

Next, we perform the TSKF method with a true model shown in Eq. 5.36 and a estimated model based on raw data, the simulation results are summarized in Table 5.3:

As shown in Table 5.3, surprisingly, applying the true model can only obtain a slightly better result than the one built on pre-whitened data and raw data. A possible explanation is that the Mahalanobis distance d_t is not sensitive enough and some changes of local variances caused by the outliers are overshadowed by local process dynamics. In other words, though some

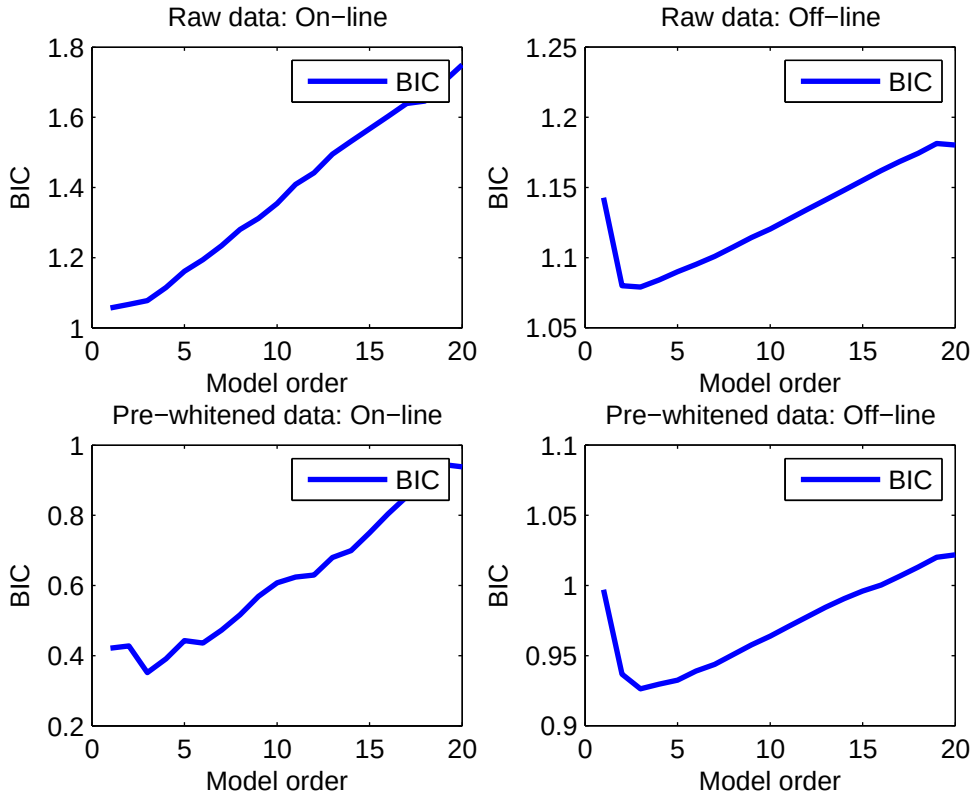


Figure 5.2: Model order selection for additive outliers

Simulation condition: ARMA (1,1), $\phi = 0.9$; $\theta = 0$; Amp=4; $\kappa = 5\%$;

Table 5.3: Model impact analysis, $\kappa = 5\%$

	on-line(*)		off-line(**)	
	$\beta(/\%)$	$\chi(/\%)$	$\beta(/\%)$	$\chi(/\%)$
True model	85.40	2.04	78.37	1.05
Pre-whitened data	82.34	2.11	75.47	1.17
Raw data	78.34	2.15	70.12	1.33

(*) Simulation condition: N=10000; rep=500; MW=100
(**) Simulation condition: N=10000; rep=500; PW=1000, $\gamma = 95\%$

increases are shown in d_t as outliers, they are not large enough to violate the threshold and raise an alarm. Furthermore, excluding the pre-whitening step from the method can negatively affect the detection results, though not significantly, because outliers affect the estimation of center μ of the data, leading to a gap showing between raw data and pre-whitened data results.

5.4.2 ARMA(1,1)model

$$\left(1 - \phi z^{-1}\right) x_t = \left(1 - \theta z^{-1}\right) \epsilon_t \tag{5.37}$$

where ϕ, θ are coefficients, ϵ_t is the white noise, $\epsilon_t \sim N(0, 1)$ and z^{-1} is a shift operator.

5.4.2.1 Additive outlier detection

The on-line and off-line additive outlier detection results for data from ARMA(1,1) processes are shown in Table 5.4:

From Table 5.4, we can see that for on-line outlier detection rate χ, the

Table 5.4: Additive outlier detection rates for data from ARMA (1, 1) process at $\kappa = 5\%$

Case No.	ϕ	θ	Δ	Amp	Liu's filter-cleaner(***)		on-line(*) TSKF		Hampel(***)		off-line(**) TSKF	
					β(/%)	χ(/%)	β(/%)	χ(/%)	β(/%)	χ(/%)	β(/%)	χ(/%)
1	0.0	0.0	2.5	3	0.82	65.13	1.51	70.12	0.72	70.91	1.42	69.03
2	0.0	0.0	2.6	4	0.55	82.83	1.07	91.78	0.63	84.23	0.50	88.17
3	0.0	0.0	2.6	5	0.43	95.41	1.12	99.15	0.44	96.41	0.11	95.76
4	0.0	-0.5	2.5	3	1.01	63.35	1.77	68.08	0.92	50.32	1.64	67.47
5	0.0	-0.5	2.7	4	0.54	78.24	1.48	89.4	0.86	73.85	0.65	85.34
6	0.0	-0.9	3.0	4	0.58	65.87	1.80	80.75	0.81	53.89	1.10	76.68
7	0.5	0.0	2.5	3	1.12	64.38	1.82	67.45	0.94	49.52	1.62	64.65
8	0.5	0.0	2.7	4	0.49	82.44	1.48	89.69	0.71	73.65	0.73	83.92
9	0.9	0.0	3.0	4	0.59	79.84	2.11	82.34	0.47	12.57	1.17	75.47

(*) Simulation condition: N=10000; rep=500; MW=100
(**) Simulation condition: N=10000; rep=500; PW=1000, $\gamma = 95\%$
(***) Some results come from Liu et al.[186].

TSKF method obtains results close to Liu's filter-cleaner[186], and both work better than the Hampel identifier when the process autocorrelation becomes high (shown in the first order correlation coefficient ϕ). The result suggests that system dynamics affect the Hampel identifier more significantly.

Moreover, a larger outlier size Amp will help the outlier detection by increasing the outlier detection rate χ, but it will not necessarily decrease the mis-identification rate β for the TSKF method.

In addition, though detecting more AOs than the Hampel identifier and on-line filter-cleaner, the TSKF method has a higher mis-identification rate β than the later two, and it increases with a higher correlation ϕ. Such a phenomenon can be attributed to that Δ is set to be a constant value and does not change as the process evolves. However, possible metrics to monitor the process changes, such as the variance and mean, will all be negatively affected by the outliers. Thus, the auto-adjustment of Δ though may lower the mis-identification rate β, will decrease the detection rate χ, as validated

by several additional simulations.

Last but not least. Though the on-line and off-line results of the TSKF method are close, it is worth mentioning that the off-line results can be obtained much faster than the on-line results since no moving window is applied.

Figs. 5.3 and 5.4 show that the TSKF method is able to detect more additive outliers in dynamic process than the Hampel identifier.

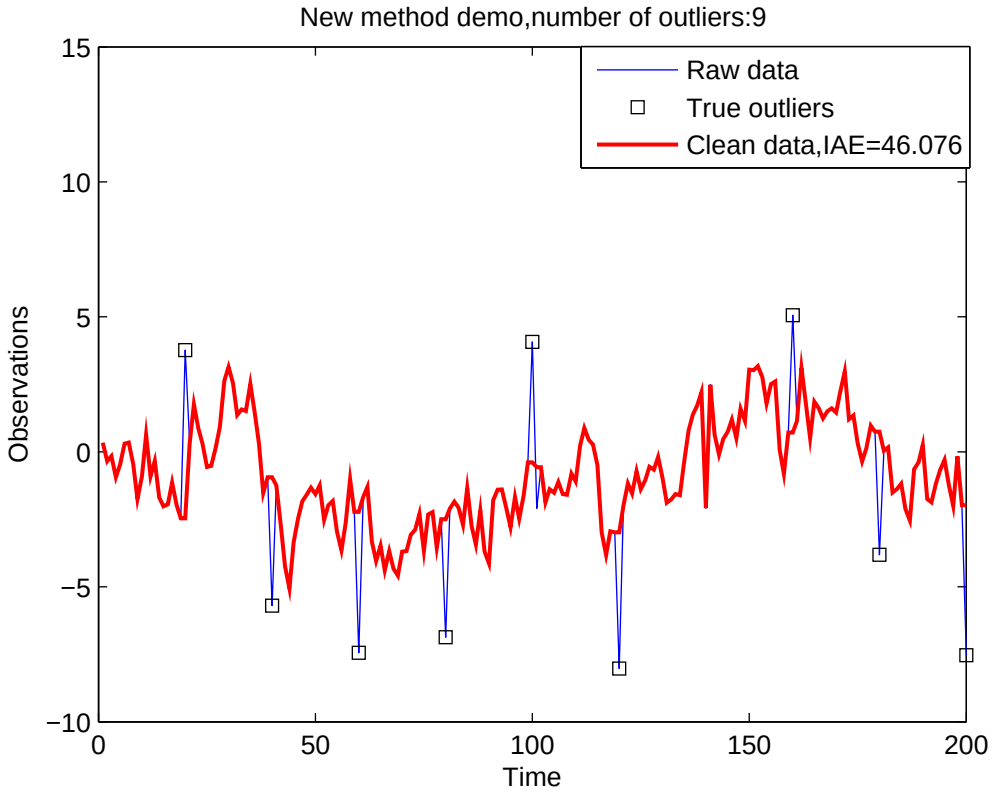


Figure 5.3: TSKF method for additive outliers

Simulation condition: ARMA (1,1), $\phi = 0.9$;
$\theta = 0$; Amp=5; $\kappa = 5\%$; MW=100

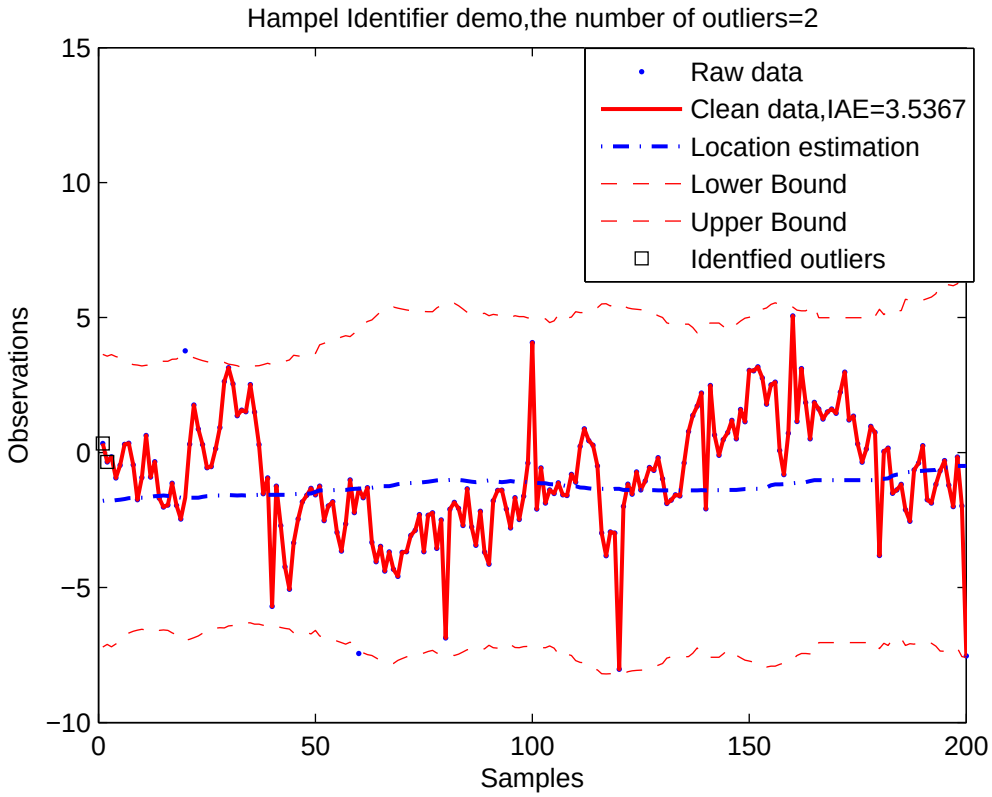


Figure 5.4: The Hampel identifier for additive outliers

Simulation condition: ARMA (1,1), $\phi = 0.9$; $\theta = 0$; Amp=5; $\kappa = 5\%$; MW=100

5.4.2.2 Innovational outlier detection

The on-line and off-line innovational outlier detection results for data from ARMA(1,1) processes are shown in Table 5.5:

Table 5.5: Innovational outlier detection results for data from ARMA (1,1) processes

Case No.	ϕ	θ	Δ	Amp	on-line(*) TSKF		on-line(*) Hampel(**)		off-line(**) TSKF	
					$\beta(/\%)$	$\chi(/\%)$	$\beta(/\%)$	$\chi(/\%)$	$\beta(/\%)$	$\chi(/\%)$
1	0.0	0.0	2.6	5	1.06	74.08	0.33	73.98	0.18	72.72
2	0.0	-0.5	2.7	5	0.84	75.54	0.45	69.02	0.17	72.58
3	0.0	-0.9	3.0	5	0.98	72.64	0.75	56.21	0.22	68.97
4	0.5	0.0	2.7	5	0.77	74.44	0.44	67.17	0.18	72.56
5	0.9	0.0	3.0	5	0.32	72.67	0.61	13.88	0.18	72.12

(*)Simulation condition: N=10000; rep=500; MW=100.
(*)Simulation condition: N=10000; rep=500; $\gamma = 99\%$

By analyzing results shown in Table 5.5, we can see that similar to the AO detection results in Table 5.4, the TSKF method works much better than the Hampel identifier in IO detection, especially when the process autocorrelation becomes high (shown in the first order correlation coefficient ϕ).

Furthermore, for the TSKF method, the detection rates of IOs do not show a lot of differences. Because unlike the AOs, the effect coming from interactions between IOs and the system dynamics, only lasts for a finite number of observations and is neutralized by a lower contamination rate(1%). This makes the IOs more difficult to detect. Even though the IOs have amplitudes

of 5, the TSKF method cannot guarantee a 100% IO detection rate χ every simulation run, and sometimes the detection rate χ is only close to 50%. Thus, the average detection rates are less than 80% and the differences are negligible.

In comparison with Fig. 5.3, Fig. 5.5 exhibits that unlike the additive outliers, the innovational outliers will impact a finite number of observations afterwards. Compared with Fig. 5.6, Fig. 5.5 also illustrates that the TSKF method is able to detect more innovational outliers than the Hampel identifier.

5.4.3 VARMA(1,1)model

$$\left(\mathbf{I} - \mathbf{\Phi z^{-1}}\right) \mathbf{x}_t = \left(\mathbf{I} - \mathbf{\Omega z^{-1}}\right) \boldsymbol{\epsilon}_t \tag{5.38}$$

where $\mathbf{\Phi}$,$\mathbf{\Omega}$ are coefficient matrices,$\boldsymbol{\epsilon}_t$ is the driving noise of the model, $\boldsymbol{\epsilon}_t \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$ and z^{-1} is a shift operator.

5.4.3.1 Additive outlier detection

The on-line and off-line additive outlier detection results for data from VARMA(1,1) processes are shown in Table 5.6:

As for parameter settings, based on additional simulations, increasing the model complexity by choosing a larger parameter l does not obtain a better performance. The parameter a is chosen to be 2 for on-line and 3 for off-line to ensure the DPCA model captures close to 90% total variance of the data set, and to prevent model over-fitting at the same time. Though increasing the significance level α will help raise the detection rate, especially during off-line

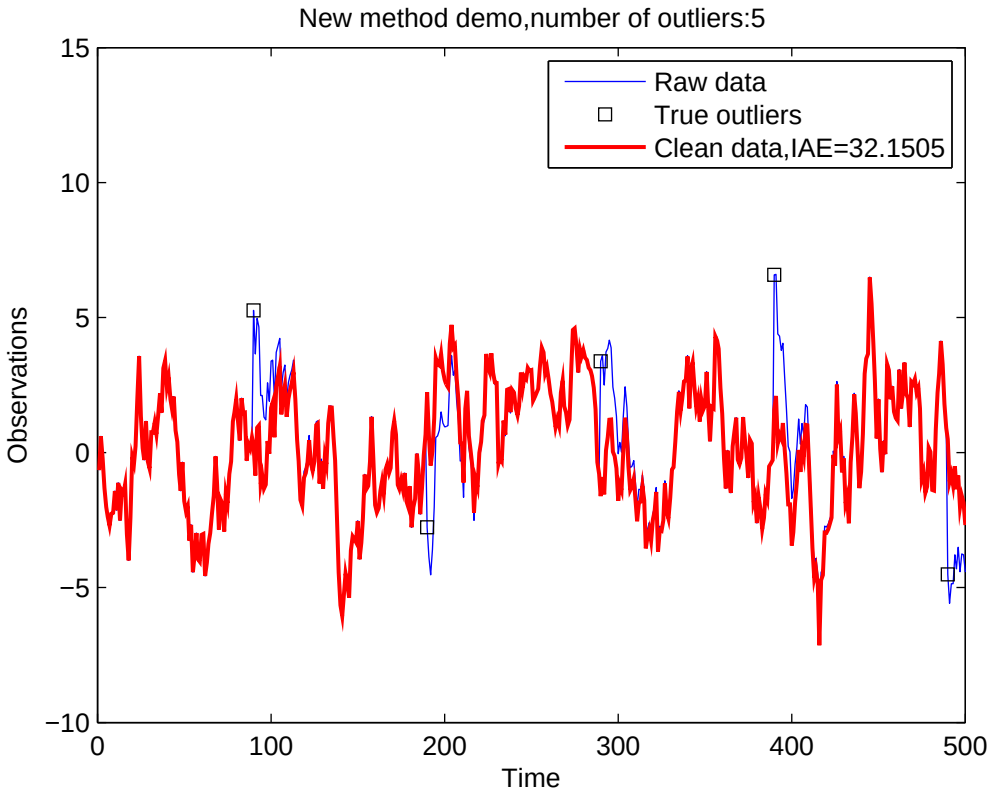


Figure 5.5: TSKF method for innovational outliers

Simulation condition: ARMA (1, 1), $\phi = 0.9$; $\theta = 0$; Amp=5; MW=100

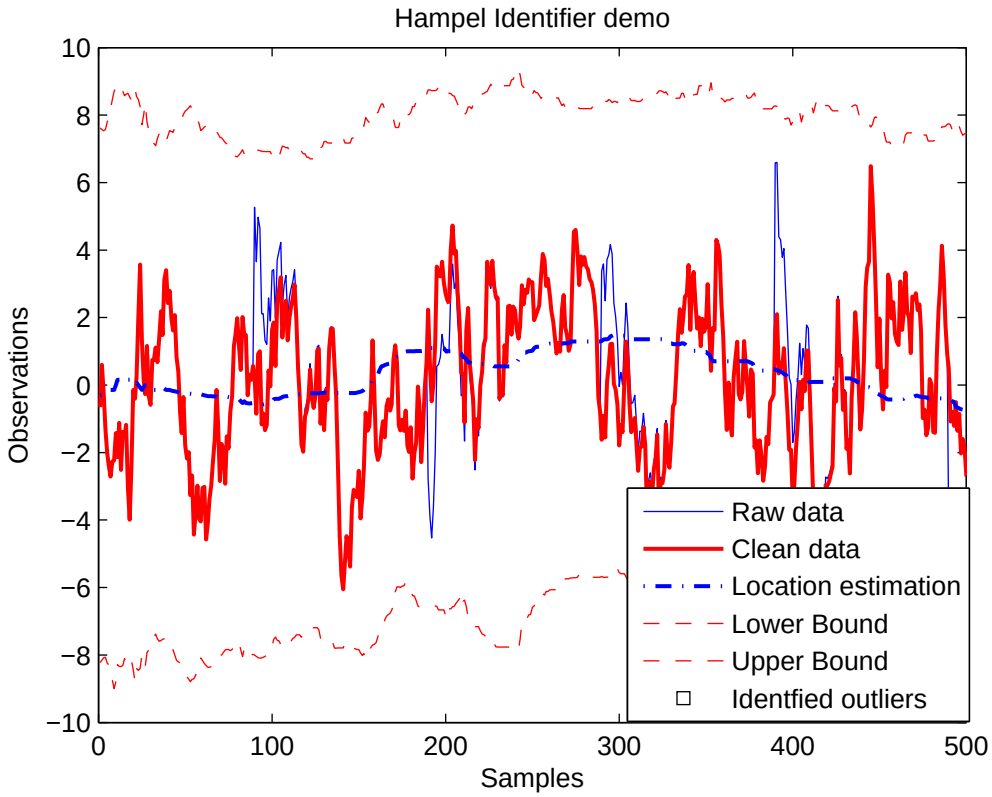


Figure 5.6: The Hampel identifier for innovational outliers

Simulation condition: ARMA (1,1), $\phi = 0.9$; $\theta = 0$; Amp=5; MW=100

Table 5.6: Additive outlier detection rate $\chi/\%$ for data from VARMA (1,1) process at $\kappa = 5\%$

Case No.	$\mathbf{\Phi}$	$\mathbf{\Omega}$	$\mathbf{\Delta}$	$\begin{bmatrix} Amp \\ Amp \end{bmatrix}$	on-line(*) TSKF	on-line(*) DPCA(I)	on-line(*) PCA(II)	off-line(**) TSKF	off-line(**) DPCA(III)	off-line(**) PCA(IV)
1	**0.0**	**0.0**	3.2	$\begin{bmatrix} 3.0 \\ 3.0 \end{bmatrix}$	90.22	71.82	72.24	88.32	45.16	66.57
2	**0.0**	**0.0**	3.2	$\begin{bmatrix} 4.0 \\ 4.0 \end{bmatrix}$	97.11	86.92	88.72	93.74	66.93	87.11
3	**0.0**	**0.0**	3.2	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	100	95.72	95.78	97.60	81.79	96.27
4	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	3.5	$\begin{bmatrix} 4.0 \\ 4.0 \end{bmatrix}$	98.22	87.12	89.88	93.69	66.74	87.39
5	$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$	3.5	$\begin{bmatrix} 4.0 \\ 4.0 \end{bmatrix}$	95.56	87.80	89.06	93.72	66.75	87.41
6	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	3.3	$\begin{bmatrix} 3.0 \\ 3.0 \end{bmatrix}$	88.47	69.02	71.42	85.21	45.24	66.33
7	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	3.3	$\begin{bmatrix} 4.0 \\ 4.0 \end{bmatrix}$	94.22	87.52	88.66	93.64	66.57	87.38
8	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	3.3	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	99.11	95.22	95.78	97.62	81.84	96.35

(*) Simulation condition: N=1000; rep=100; MW=100.
(**) Simulation condition: N=10000; rep=100; PW=1000; $\gamma = 95\%$.
(I) Parameter setting: $\alpha = 0.01$; $l = 1$; $a = 2$.
(II) Parameter setting: $\alpha = 0.01$; $a = 1$.
(III) Parameter setting: $\alpha = 0.01$; $l = 1$; $a = 3$.
(IV) Parameter setting: $\alpha = 0.01$; $a = 1$.

implementation, it will significantly affect the results by leading to a much larger mis-identification rate β (larger than 10%). Thus, α was chosen to be 0.01.

By analyzing results shown in Table 5.6 we can see that the TSKF method generally obtains a higher outlier detection rate χ than the DPCA and PCA method. Surprisingly, the DPCA method does not obtain as good detection rates as PCA, especially during off-line implementation. Such a result can be largely attributed to that without a pre-whitening step, an outlier contamination rate of 5% will severely damage the serial correlation in augmented matrix shown in Eq. (5.8) and lead to inaccurate t^2 results.

Moreover, a larger outlier size *Amp* will help the outlier detection by increasing the outlier detection rate χ, while increasing autocorrelation $\mathbf{\Phi}$ will

negatively affect the outlier detection results.

Fig. 5.7 shows the on-line multivariate additive outlier detection results of the TSKF method for a VARMA(1,1) process. Fig. 5.8 shows the Hotelling's T^2 record on the first moving window of the DPCA method, it needs to be pointed that though time points 40 and 80 are missed, they can be successfully detected in moving windows afterwards.

5.4.3.2 Innovational outlier detection

The on-line and off-line innovational outlier detection results for data from VARMA(1,1) processes are shown in Tables 5.7:

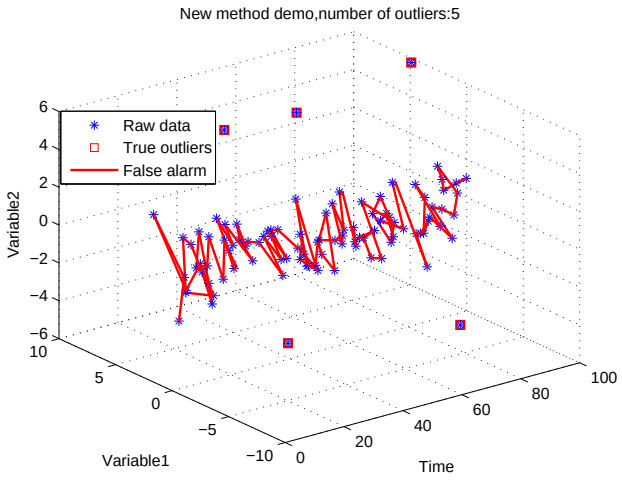
Table 5.7: Innovational outlier detection rate $\chi\%$ for data from VARMA (1,1) processes

Case No.	Φ	Ω	Δ	$\begin{bmatrix} Amp \\ Amp \end{bmatrix}$	on-line(*) TSKF	on-line(*) DPCA(I)	on-line(*) PCA(II)	off-line(**) TSKF	off-line(**) DPCA(III)	off-line(**) PCA(IV)
1	**0.0**	**0.0**	3.2	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	74.07	75.00	75.10	74.99	75.25	74.14
2	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	3.5	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	77.78	75.40	74.70	75.13	74.99	75.04
3	$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$	3.5	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	79.33	76.80	76.0	74.44	75.02	74.73
4	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$	3.3	$\begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$	75.11	76.30	76.40	75.58	74.28	74.84

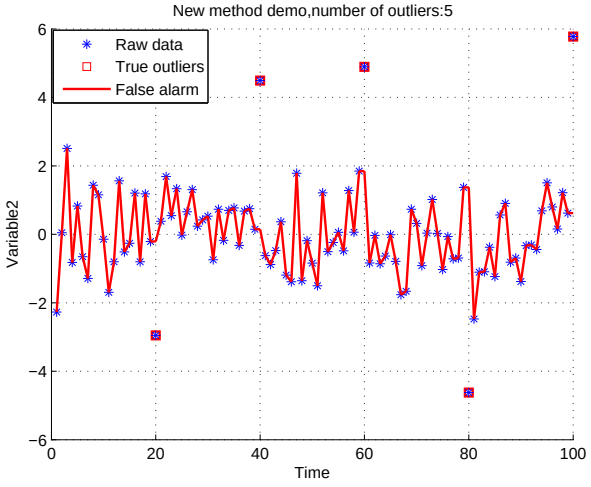
(*) Simulation condition: N=1000; rep=100; MW=100.
(**) Simulation condition: N=10000; rep=100; $\gamma = 99\%$.
(I) Parameter setting: $\alpha = 0.01$; $l = 1$; $a = 3$.
(II) Parameter setting: $\alpha = 0.01$; $a = 1$.
(III) Parameter setting: $\alpha = 0.01$; $l = 1$; $a = 3$.
(IV) Parameter setting: $\alpha = 0.01$; $a = 1$.

It is found that the DPCA with a significance level $\alpha = 0.01$ obtains desirable results for both on-line and off-line cases. In addition, the parameter a is chosen to be 3 and l to be 1 for the same reason discussed in univariate outlier cases.

Furthermore, for the same reasons as discussed in ARMA(1,1) cases,



(a) 3D view



(b) Front view

Figure 5.7: Additive outlier detection results obtained by the TSKF method

Simulation condition: VARMA(1,1); on-line; case No.=5

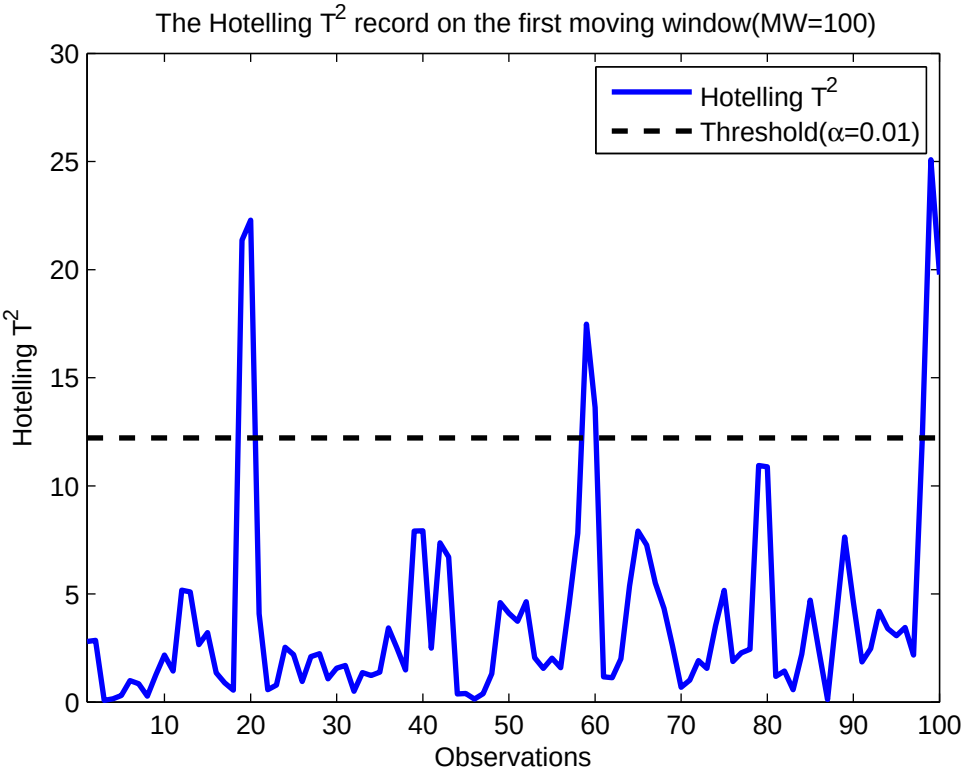


Figure 5.8: Hotelling's T^2 record on the first moving window of dynamic PCA

Simulation condition: VARMA(1,1); on-line; $\kappa = 5\%$; case No.=5; additive outlier

IO detection results of TSKF, PCA and DPCA in Table 5.7 are close and close to 75%. It is worth mentioning that the PCA and DPCA methods are faster than the TSKF method in obtaining the results.

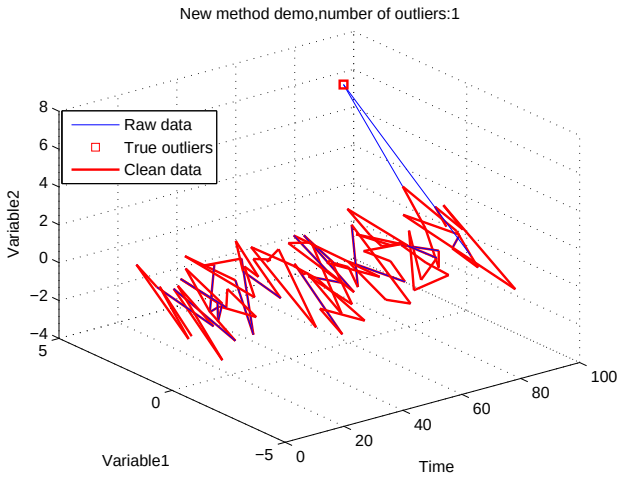
Fig. 5.9 and Fig. 5.10 show the multivariate innovational outlier detection results of the TSKF method and the DPCA method for a VARMA(1,1) process, respectively. Comparing these two figures, we can see both methods correctly identified the location of pulse function in the first moving window.

5.4.4 Summary and discussion of simulation testing results

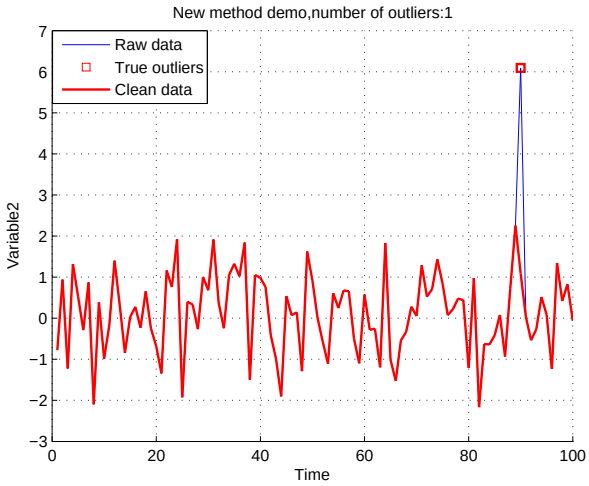
Based on the simulation results, we can see that while the additive outliers only affect single observations, innovational outliers will have an impact on a finite number of observations after they appear. A larger autocorrelation contained in the process data will negatively affect the detection results, while a larger outlier size will help the outlier detection.

Furthermore, though the interactions between IOs and the system dynamics make them more difficult to be detected than the additive outliers, a desirable detection rate χ can still be obtained if the contamination rate $\kappa$is low and the amplitude of IOs is high.

Though only the outlier detection results of ARMA(1,1) and VARMA(1,1) processes are shown, the TSKF method works well for higher order stationary process data, such as ARMA(2,1) and VARMA(2,1). The TSKF method works better than the Hampel identifier, PCA and DPCA methods in detection of AOs. For IOs, the Hampel identifier becomes incompetent in detecting



(a) 3D view



(b) Front view

Figure 5.9: Innovational outlier detection results obtained by the TSKF method

Simulation condition: VARMA(1,1); on-line; case No.=3; N=1000

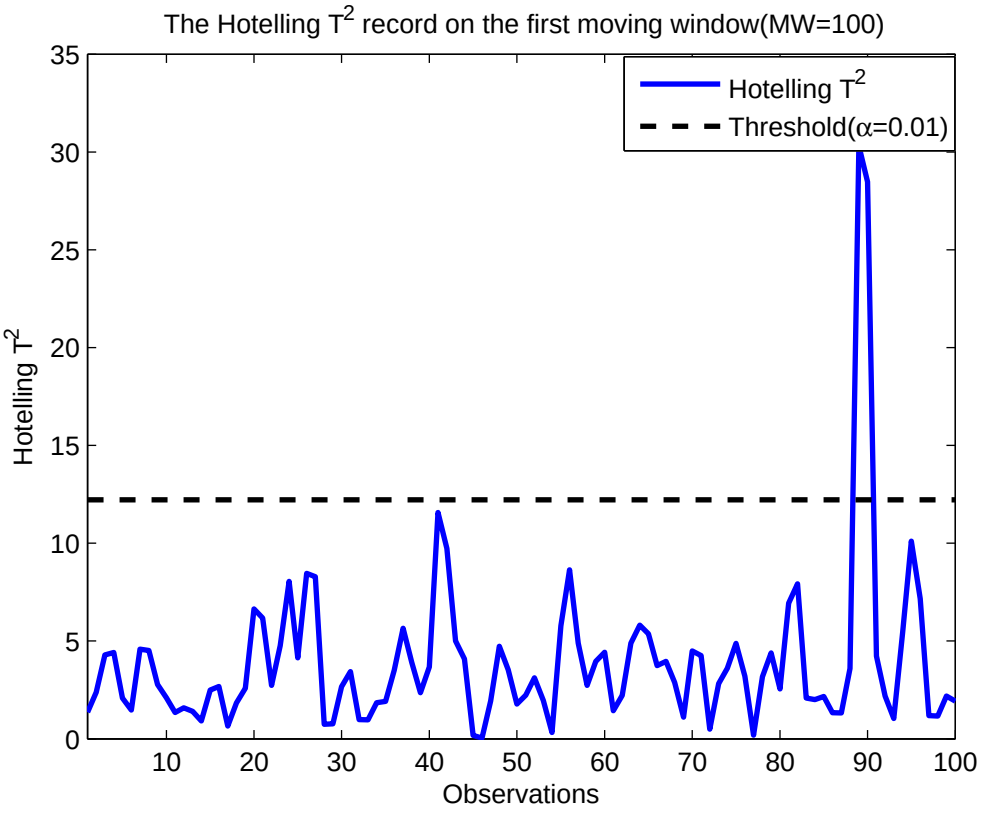


Figure 5.10: Hotelling's T^2 record on the first moving window of dynamic PCA

Simulation condition: VARMA(1,1); on-line; case No.=3; N=1000; innovational outlier

them when a high autocorrelation exists. The TSKF, PCA and DPCA still obtain close results in detecting IOs, but the computational cost of the TSKF method is higher.

5.5 Plant data Testing

In this section, an industrial data set obtained from a chemical plant is used to perform on-line testing of the TSKF method.

Two variables shown in Fig. 5.11 have been selected from the process data set because outliers and dynamic changes are present in both variables. The missing values have been replaced with local means. While the second variable shows many step changes, the first variable contains dynamic responses caused by a series of step changes.

The model order selection is based on BIC for two variables, as shown in Fig. 5.12. Although Var1 and Var2 vary with time, a model order of 1 or 2 suffice. In this case, we pick 2.

The on-line outlier detection process is carried out as follows: each variable is cleaned alone by the TSKF method and the Hampel identifier. The results are plotted in Figures 5.13 to 5.16. Compare Figure 5.13 with 5.14, we can see that the TSKF method is able to detect more spikes between time points 200 and 300 than the Hampel identifier, and successfully replaces an outlier shown at time point 700. Similar results can also be found near time point 700 in Figure 5.15 and 5.16.

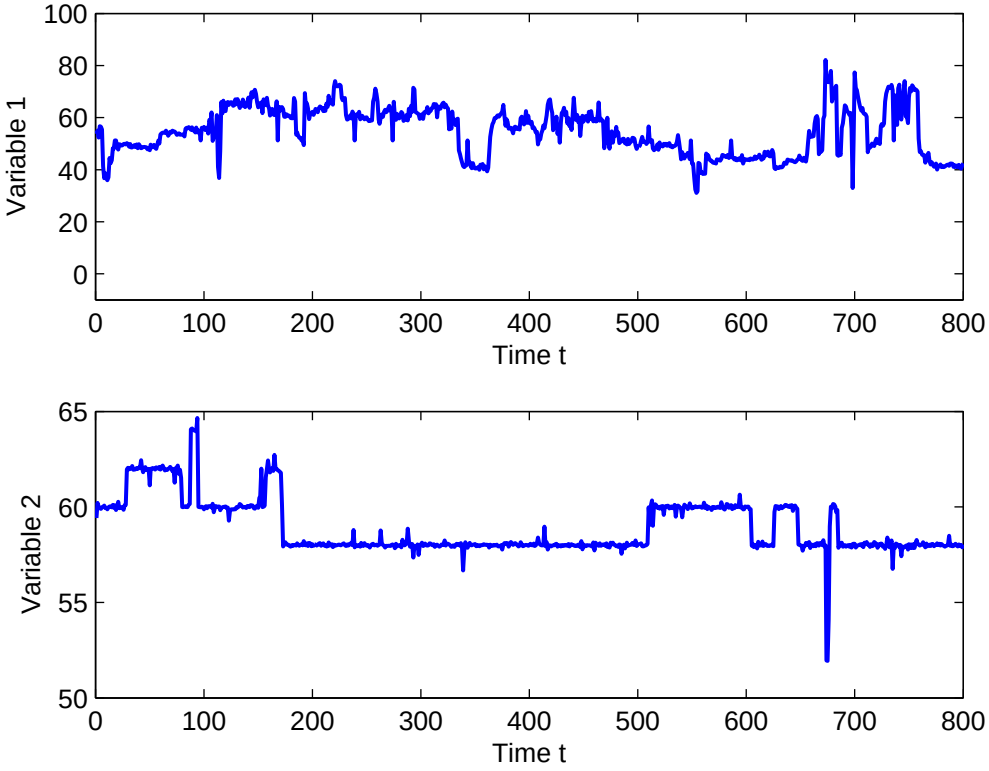


Figure 5.11: Raw plant data

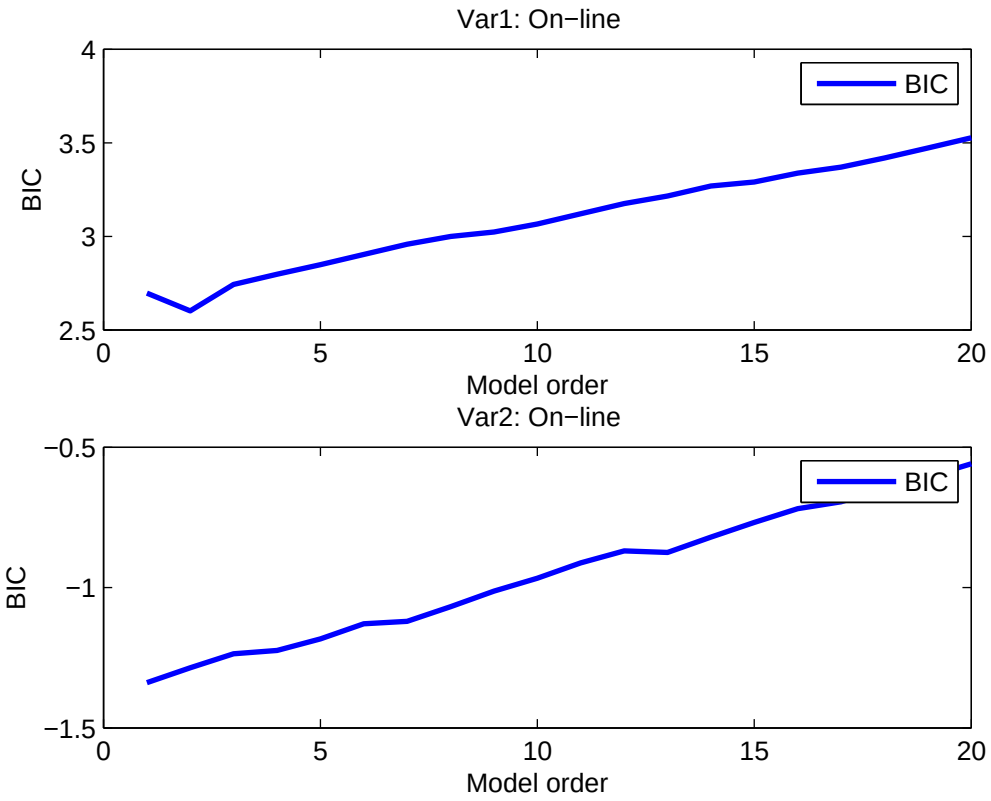


Figure 5.12: BIC of raw plant data at single moving window

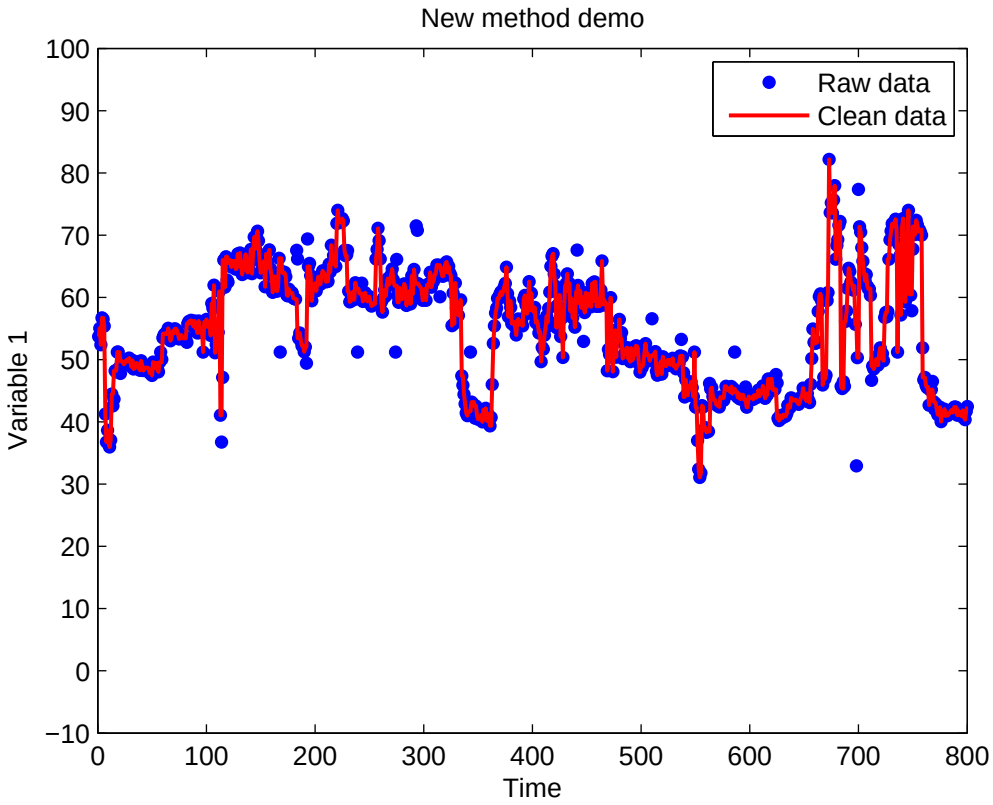


Figure 5.13: The TSKF method for V1

Simulation condition: $\Delta = 5$; MW=10; on-line testing

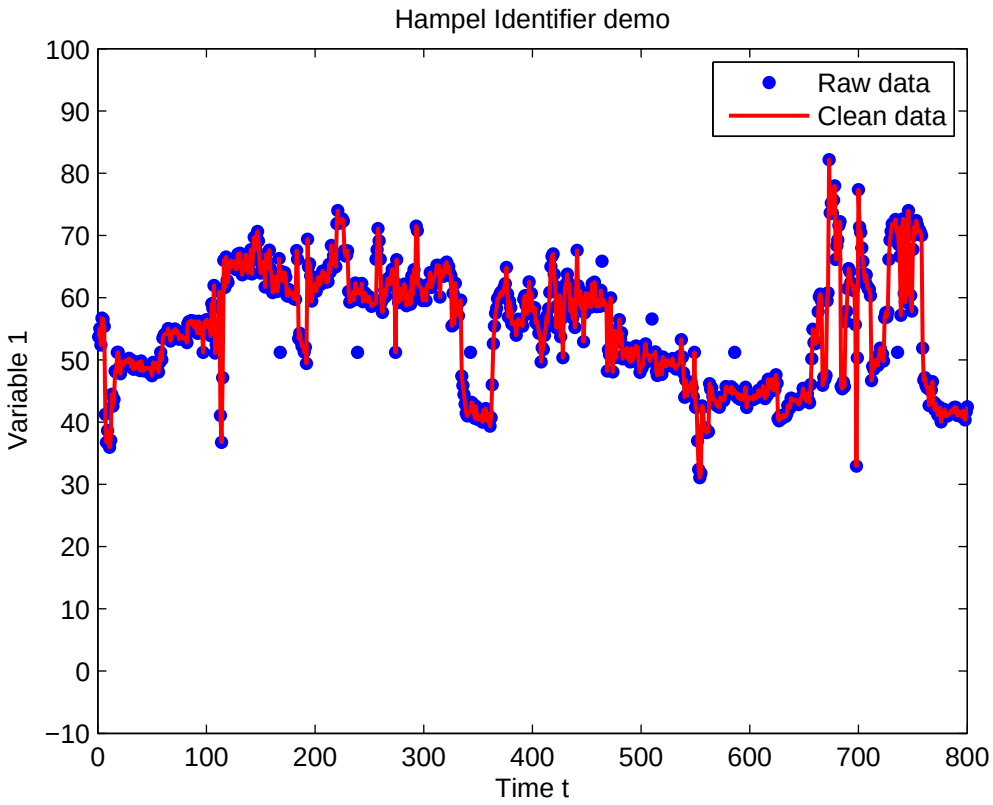


Figure 5.14: The Hampel identifier for V1

Simulation condition: MW=10; on-line testing

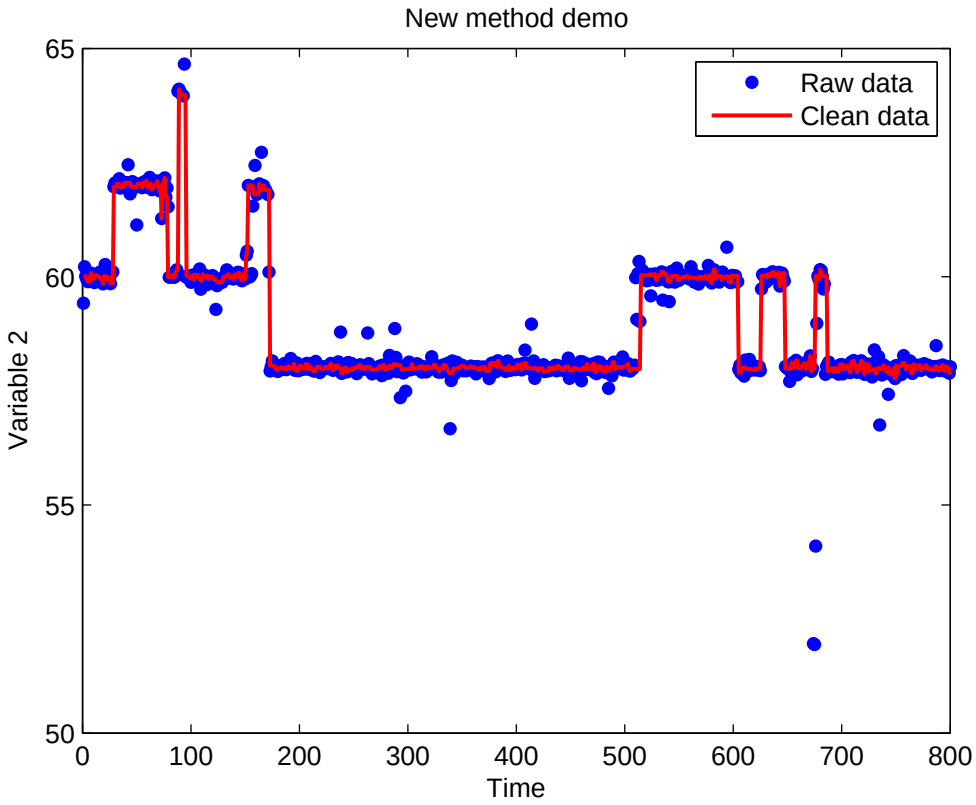


Figure 5.15: The TSKF method for V2

Simulation condition: $\Delta = 1.5$; MW=10; on-line testing

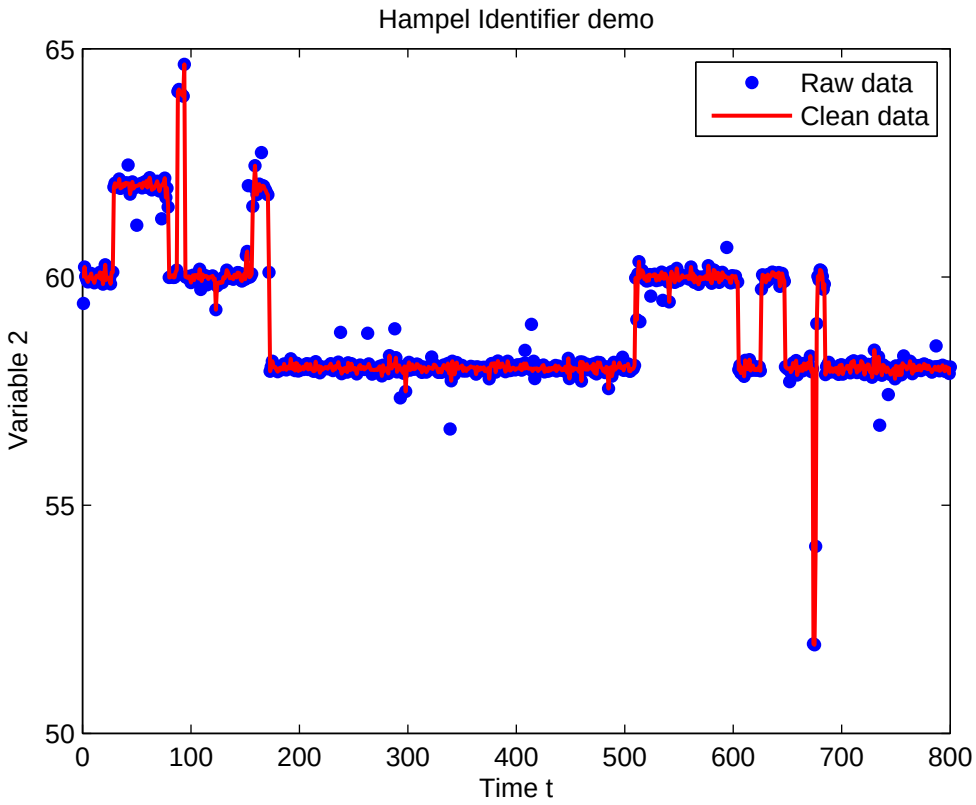


Figure 5.16: The Hampel identifier for V2

Simulation condition: MW=10; on-line testing

5.6 Summary

In this chapter, a new method (TSKF) for outlier detection has been proposed that is suitable for both univariate and multivariate outlier detection in dynamic data sets. Both on-line and off-line versions and related parameter tuning for the new method have been given.

Different from the on-line filter-cleaner[186], the TSKF method incorporates the Burg-type time series model fitting algorithm, which ensures stability of the method when dealing with ill-conditioned auto-covariance matrices in multivariate cases. In addition, neighboring normal points will be used to replace outliers instead of using imposed model predictive value. Moreover, obtaining model parameters directly from the preliminary clean data set has a lower computational complexity in comparison with the procedure used in on-line filter-cleaner (converting the original data to more complicated matrices and applying reweighted MCD method to estimate parameters).

Based on the simulation testing results, the TSKF method outperforms the Hampel identifier and the dynamic PCA method in additive outlier detection in an ARMA(1,1) process and a VARMA(1,1) process, respectively. Interestingly, for IOs, the TSKF method though outperforms the Hampel identifier in an ARMA(1,1) process, does not differ a lot from PCA and DPCA in a VARMA(1,1) process, due to combining effects of interactions between IOs and system dynamics, as well as contamination rate and outlier amplitude, as discussed in simulation section. It is worth mentioning that the computational cost of the TSKF method is higher than the PCA and DPCA method, which

makes it a less competitive choice.

Based on actual plant data testing results, the TSKF method is able to detect more univariate outliers than the Hampel identifier.

Last but not least, tests have been made on non-stationary processes, in which the IOs will lead to permanent parameter drifts and shifts, and the results of TSKF on detecting those changes are not desirable. Thus, the method still needs to be improved on dealing with such problems in the future work.

Chapter 6

Study of information transfer in the frequency domain

6.1 Motivation

As discussed in Chapter 1, modern chemical plant is massively instrumented with IP-enabled intelligent sensors, controllers, and actuators collecting on-line process information. Such enormous amounts of data facilitate the development of data-driven methods such as principal component analysis (PCA)[168, 195, 64] that provide process engineers with more convenient and intuitive ways to monitor the process than traditional first-principles models. One of the most important issues in process monitoring is to detect and diagnose process disturbances such as plant-wide oscillations [292]. Plant-wide oscillations can occur largely due to poorly tuned controllers, actuator nonlinearities, and possible external oscillatory disturbances [90, 291, 335]. Because the material and information flow streams between units are highly correlated, oscillations generated at one point will propagate to the unit or even the whole plant through connecting flow streams and may cause poor control performance, inferior quality products, excessive energy consumption, and even plant safety issues [292]. A distinguishing feature of oscillations is that, unlike other disturbances, they have similar spectral patterns in affected variables

and can be detected using power spectra. A spectral envelope method used by Jiang et al. (2007) [149] provides an intuitive and fast way to find the shared oscillation frequency and related contributing variables so that power spectra on individual variables are no longer needed. After detecting the oscillations, the next step is to determine the root causes of such oscillations from contributing variables. Generally, methods for root cause diagnosis of plant-wide oscillations can be categorized into process data-based analysis methods and topology-based methods [90]. While the process data-based methods such as the oscillation contribution index (OCI) [149] aim to analyze nonlinearity and power spectrum features, the topology-based methods are either based on process flow sheets and qualitative models [150, 331] or based on a nominal map exhibiting causal relationships. There are two types of methods to construct such a casual map: time-domain based ones including cross-correlation analysis [115], nearest neighbors [28], Granger causality [117, 14, 335] and transfer entropy [27, 91]; frequency-domain based ones including directed transfer functions [156], partial directed coherence [20], and spectral granger causality [25, 335]. Analyzing data within the frequency domain helps the users focus on frequencies containing useful information and getting rid of unwanted noise. In this paper, we will derive a spectral transfer entropy expression in Section 6.2 for root cause diagnosis based on spectral Granger causality. Brief descriptions of the spectral envelope method and related oscillation contribution index (OCI) [149] are also included in that section. In Section 6.3, through simulated and industrial case studies, the effectiveness of the new information

transfer method is demonstrated, and the diagnosis result is compared with the oscillation contribution index (OCI) method [149]. To facilitate the visualization of oscillation frequencies, a wavelet power spectrum of the root cause variable is provided.

6.2 Methods description

6.2.1 Spectral envelope method

First proposed by Stoffer et al. (1993) [284], the spectral envelope method has been applied to detect common signals in a series of time series [283, 285], and successfully pinpointed the plant-wide oscillations with common frequencies in industrial applications [149].

Assume $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_n(t)]^T$ to be a time series on $\Re^n$, and each variable $x_i(t)$ has been auto-scaled to have identical power. Define the spectral envelope to be:

$$\lambda(\omega) = \sup_{\beta \neq \mathbf{0}} \left\{ \frac{\beta^* \mathbf{P_x}(\omega) \beta}{\beta^* \mathbf{V} \beta} \right\} \tag{6.1}$$

where β is an n-dimensional real or complex vector, * represents the conjugate transpose, $\mathbf{V_x}$ and $\mathbf{P_x}(\omega)$ denote the covariance matrix and the power spectral density (PSD) matrix of $\mathbf{x}$ respectively, and ω is the normalized frequency satisfying $-1/2 \leq \omega \leq 1/2$.

$\lambda(\omega)$ in Eq. (6.1) denotes the largest portion of power that can be obtained at frequency ω. Since the data have been normalized, $\lambda(\omega)$ is the

largest eigenvalue of $\mathbf{P}_{\mathbf{x}}(\omega)$, and $\beta(\omega)$ is the corresponding eigenvector. Assume $\mathbf{x}(t)$ has N samples, i.e., $t = 0, 1, ..., N-1$. The Fourier frequencies are defined as $\omega_k = k/N$, for $k = 1, 2, ..., [N/2]$, where $[N/2]$ is the greatest integer less than or equal to $N/2$. A simple estimate of $\mathbf{P}_{\mathbf{x}}(\omega_k)$ is given by:

$$\hat{\mathbf{P}}_{\mathbf{x}}(\omega_k) = \sum_{j=-r}^{r} h_j \hat{\mathbf{I}}_N(\omega_{k+j}) \tag{6.2}$$

where h_j is symmetric positive weights satisfying $h_j = h_{-j}$ ($\{h_0 = 3/9, h_{\pm 1} = 2/9, h_{\pm 2} = 1/9\}$ in this paper), and

$$\hat{\mathbf{I}}_N(\omega_k) = \frac{1}{N}\left[\sum_{t=0}^{N-1} \mathbf{x}(t) \exp(-2\pi i t \omega_k)\right]\left[\sum_{t=0}^{N-1} \mathbf{x}(t) \exp(-2\pi i t \omega_k)\right]^* \tag{6.3}$$

Assume that $\lambda_1(\omega) = \lambda(\omega), \lambda_2(\omega), ..., \lambda_n(\omega)$ are the eigenvalues of $\hat{\mathbf{P}}_{\mathbf{x}}(\omega)$ arranged in decreasing order, and $\beta_1(\omega) = \beta(\omega), \beta_2(\omega), ..., \beta_n(\omega)$ are the corresponding eigenvectors. The asymptotic covariance matrix of the sample optimal scaling vector $\hat{\beta}(\omega)$ is given by

$$\mathbf{V}_{\beta}(\omega) = v^{-2}\lambda_1(\omega) \sum_{l=2}^{n} \lambda_l(\omega) [\lambda_1(\omega) - \lambda_l(\omega)]^{-2} \beta_l(\omega) \beta_l^*(\omega) \tag{6.4}$$

where $v = \left(\sum_{j=-r}^{r} h_j^2\right)^{-\frac{1}{2}}$. The distribution of $2\left|\hat{\beta}_j(\omega) - \beta_j(\omega)\right|^2 / \sigma_j(\omega)$ approximately follows a Chi-square distribution, where $\sigma_j(\omega)$ is the jth diagonal element of $\mathbf{V}_{\beta}(\omega)$, and $\hat{\beta}_j(\omega)$ and $\beta_j(\omega), j = 1, ..., n$ are the jth element of the estimated and true optimal scaling vector respectively. If $2\left|\hat{\beta}_j(\omega)\right|^2 / \sigma_j(\omega)$ violates the threshold $\chi_2^2(\alpha)$, we assume oscillation at frequency ω is contained because the null hypothesis $\beta_j(\omega) = 0$ is rejected with $(1-\alpha)$ confidence.

6.2.2 Oscillation contribution index

The oscillation contribution index of $x_j(t)$ is defined to be:

$$OCI_j(\omega) = \frac{\left|\hat{\beta}_j(\omega)\right|}{2\sigma_{\hat{\beta}}(\omega)} \tag{6.5}$$

where $\sigma_{\hat{\beta}}(\omega)$ is the standard deviation of the magnitude of the optimal scalings of all the oscillating variables. Commonly, variables having $OCI(\omega) > 1$ are considered as the root cause variables at frequency ω with the larger contributions at the spectral envelope peak [90].

6.2.3 Spectral Granger causality

Granger causality [117], originally developed within the context of econometric theory, is broadly applied in determining the "predictive causality "between two time series in areas such as neurosciences [25, 286] and process modeling [335]. A powerful feature of Granger causality is that it can be decomposed by frequency [113] so that spectral causality at specific frequencies can be calculated, which is very useful when the signals contain unwanted frequencies. Assume $X(t)$ and $Y(t)$ are two time series describing stationary stochastic processes, a bivariate autoregressive (AR) model can be built as follows:

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=1}^{p} \begin{bmatrix} a_{xx,k} & a_{xy,k} \\ a_{yx,k} & a_{yy,k} \end{bmatrix} \begin{bmatrix} X(t-k) \\ Y(t-k) \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix} \tag{6.6}$$

where $a_{ij,k}$ are the AR coefficients, and the residual covariance matrix as:

$$\Sigma = \operatorname{cov} \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{6.7}$$

If we perform Fourier transform on Eq. (6.6), we obtain:

$$\begin{bmatrix} A_{xx}(\omega) & A_{xy}(\omega) \\ A_{yx}(\omega) & A_{yy}(\omega) \end{bmatrix} \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} E_x(\omega) \\ E_y(\omega) \end{bmatrix} \tag{6.8}$$

where the components of the coefficient matrix $[A_{ij}(\omega)]$ are $A_{ij}(\omega) = \delta_{ij} - \sum_{k=1}^{p} a_{ij,k} e^{-i\omega k}$. Denote the transfer function matrix $H(\omega) = [A_{ij}(\omega)]^{-1}$, and Eq. (6.8) can be re-written as follows:

$$\begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} H_{xx}(\omega) & H_{xy}(\omega) \\ H_{yx}(\omega) & H_{yy}(\omega) \end{bmatrix} \begin{bmatrix} E_x(\omega) \\ E_y(\omega) \end{bmatrix} \tag{6.9}$$

Thus, the spectral density matrix $S(\omega)$ is derived as:

$$S(\omega) = \langle X(\omega) X^*(\omega) \rangle = H(\omega) \Sigma H^*(\omega) \tag{6.10}$$

where Σ is the covariance of the full model residuals and $X^*(\omega)$ is the Hermit transpose of $X(\omega)$. For the X_1 process, pre-multiplying Eq. (6.9) on both sides with $\begin{bmatrix} 1 & 0 \\ -\Sigma_{12}/\Sigma_{11} & 1 \end{bmatrix}$, we obtain:

$$\begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} \tilde{H}_{xx}(\omega) & \tilde{H}_{xy}(\omega) \\ \tilde{H}_{yx}(\omega) & \tilde{H}_{yy}(\omega) \end{bmatrix} \begin{bmatrix} E_x(\omega) \\ \tilde{E}_y(\omega) \end{bmatrix} \tag{6.11}$$

where

$$\tilde{E}_y(\omega) = E_y(\omega) - \frac{\Sigma_{xy}}{\Sigma_{xx}} E_x(\omega) \tag{6.12}$$

and

$$\begin{bmatrix} \tilde{H}_{xx}(\omega) & \tilde{H}_{xy}(\omega) \\ \tilde{H}_{yx}(\omega) & \tilde{H}_{yy}(\omega) \end{bmatrix} = \begin{bmatrix} H_{xx}(\omega) + \frac{\Sigma_{xy}}{\Sigma_{xx}} H_{xy}(\omega) & H_{xy}(\omega) \\ H_{yx}(\omega) + \frac{\Sigma_{xy}}{\Sigma_{xx}} H_{xx}(\omega) & H_{yy}(\omega) \end{bmatrix} \tag{6.13}$$

The spectrum of X is represented by an "intrinsic"term and a "causal"term:

$$S_{xx}(\omega) = \tilde{H}_{xx}(\omega) \Sigma_{xx} H^*_{xx}(\omega) + \tilde{H}_{xy}(\omega) \tilde{\Sigma}_{yy} \tilde{H}^*_{xy}(\omega) \tag{6.14}$$

where $\tilde{\Sigma}_{yy} = \Sigma_{yy} - \left(\Sigma_{xy}^2/\Sigma_{xx}\right)$. Therefore, the spectral Granger causality from Y to X at frequency f is:

$$I_{Y\to X}(\omega) = \ln\left(\frac{|S_X|}{\left|S_{X|Y}\right|}\right) = \ln\left(\frac{|S_{xx}(\omega)|}{\left|S_{xx}(\omega) - \tilde{H}_{12}(\omega)\,\tilde{\Sigma}_{22}\tilde{H}_{12}^{*}(\omega)\right|}\right) \tag{6.15}$$

6.2.4 Spectral transfer entropy

The concept of transfer entropy was first proposed by Schreiber (2000), which measures the amount of directed transform of information between two random variables [269], and it has been applied in process industries for diagnosis of root cause of oscillation [27, 91, 90]. In the time domain, the transfer entropy between two variables is defined as:

$$T_{Y\to X} = H\left(X|X^{-}\right) - H\left(X|X^{-}\oplus Y^{-}\right) \tag{6.16}$$

where $H(X)$ is the Shannon entropy of X, $H(\cdot|\cdot)$ is the conditional entropy, $\oplus$ denotes a joint relation, and X^{-}, Y^{-} stand for the past values of X, Y. Given the past values of X, the transfer entropy can be seen as the reduction of uncertainty in the future values of X by knowing the past values of Y. In the frequency domain, we define the spectral transfer entropy in a similar fashion:

$$T_{Y\to X} = H\left(X(\omega)\right) - H\left(X(\omega)\,|Y(\omega)\right) \tag{6.17}$$

since the Fourier transform of a Gaussian is also a Gaussian (Proof see Appendix C). For Gaussian variables, we have [24]:

$$H(X) = \frac{1}{2}\ln\left(|\Sigma(X)|\right) + \frac{1}{2}n\ln(2\pi e) \Rightarrow \frac{1}{2}\ln\left(|S_X|\right) + \frac{1}{2}n\ln(2\pi e) \tag{6.18}$$

$$H\left(X|Y\right)=\frac{1}{2}\ln\left(\left|\Sigma\left(X|Y\right)\right|\right)+\frac{1}{2}n\ln\left(2\pi e\right)\Rightarrow\frac{1}{2}\ln\left(\left|S_{X|Y}\right|\right)+\frac{1}{2}n\ln\left(2\pi e\right) \tag{6.19}$$

From Eqs. 6.18 and 6.19, we have:

$$T_{Y\to X}=H\left(X\left(\omega\right)\right)-H\left(X\left(\omega\right)|Y\left(\omega\right)\right)=\frac{1}{2}\ln\left(\frac{\left|S_X\right|}{\left|S_{X|Y}\right|}\right)=\frac{1}{2}I_{Y\to X}\left(\omega\right) \tag{6.20}$$

Thus, in the frequency domain, for Gaussian variables, the Granger causality and transfer entropy are linear correlated by a factor of 2. Thus, the spectral entropy can be calculated in the same ways as spectral Granger causality, including AR model identification and Fourier transformation [25]. A similar result was obtained in the time domain by Barnett et al. (2009)[24].

6.3 Case study

In this section, the new information transfer method including the spectral envelope and the spectral transfer entropy calculation will be tested based on a simulated and an industrial data set.

6.3.1 Simulated data

Five time series are generated by an autoregressive (AR) model shown in Eq. (6.21)[20]:

$$\begin{aligned}
X_{1,n}&=0.95\sqrt{2}X_{1,n-1}-0.9025X_{1,n-2}+w_{1,n}\\
X_{2,n}&=0.5X_{1,n-2}+w_{2,n}\\
X_{3,n}&=-0.4X_{1,n-3}+w_{3,n}\\
X_{4,n}&=-0.5X_{1,n-2}+0.25\sqrt{2}X_{4,n-1}+0.25\sqrt{2}X_{5,n-1}+w_{4,n}\\
X_{5,n}&=-0.25\sqrt{2}X_{4,n-1}+0.25\sqrt{2}X_{5,n-1}+w_{5,n}
\end{aligned} \tag{6.21}$$

where $w_{1,n}, w_{2,n}, w_{3,n}, w_{4,n}, w_{5,n}$ are drawn from Gaussian noise with zero mean and unit variance. As shown in the model, it is clear that X_1 is the root cause of the oscillation. The independent power spectrum shown in Fig. 6.1 suggests that the variables share a common oscillation at 0.12 Hz. If we apply the spectral envelope method on the simulated data set, the dominant frequency is clearly revealed, as shown in Fig. 6.2. To find out the root cause of the oscillation, the oscillation contribution index (OCI) was calculated in Table 6.1. The OCI of X_1 has a larger value than others, and it suggests that X_1 is likely to be the source.

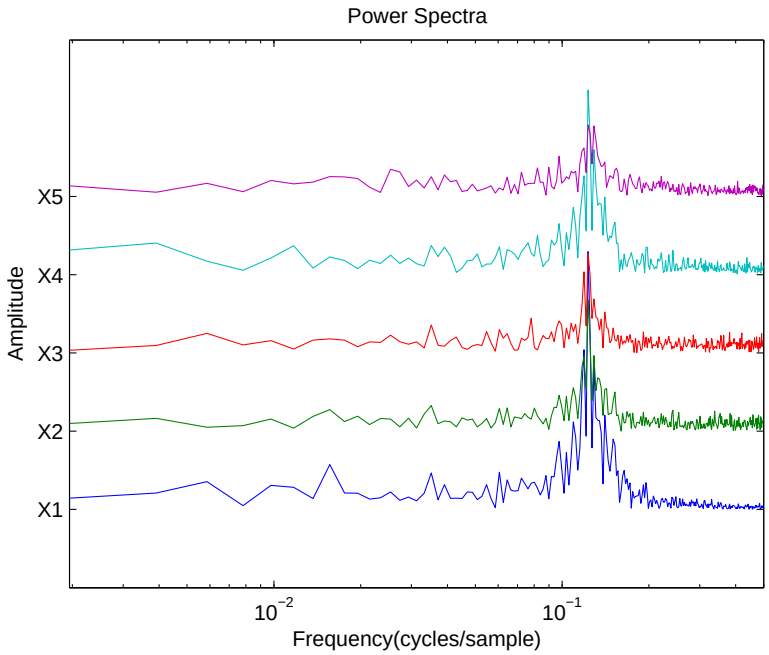


Figure 6.1: Power spectrum of the simulated data

The spectral transfer entropy matrix at $f = 0.12Hz$ is shown in Fig.

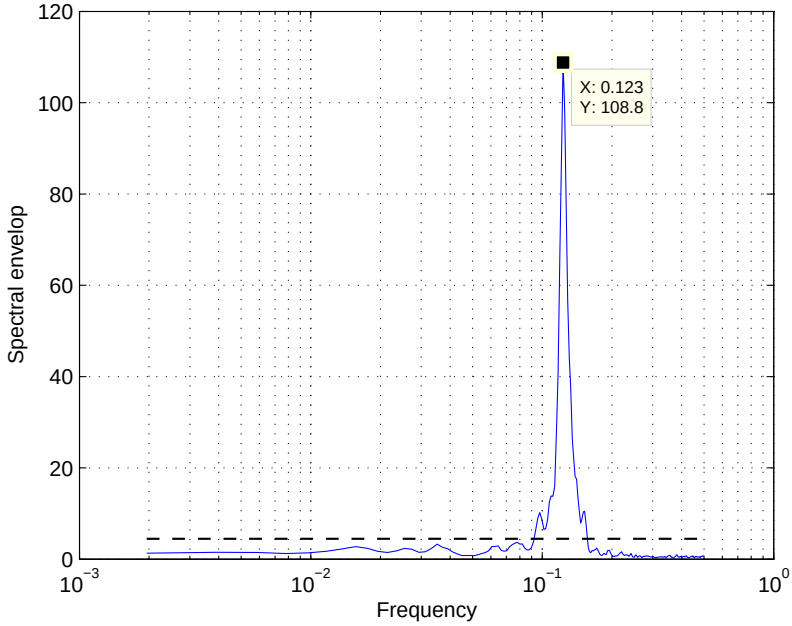


Figure 6.2: Spectral envelope of the simulated data

Table 6.1: Oscillation contribution index

Variable	OCI
X_1	1.6247
X_2	0.0408
X_3	0.9952
X_4	0.1960
X_5	1.0071

6.3, and the darkness is proportional to the strength of causality. X_1 strongly affects X_2 and X_4, and also exerts moderate causal effects on X_3. Moreover, there is a mutual causal relationship, though not very strong, between X_4 and X_5. The process topology diagram was drawn based on the transfer entropy matrix, as shown in Fig. 6.4:

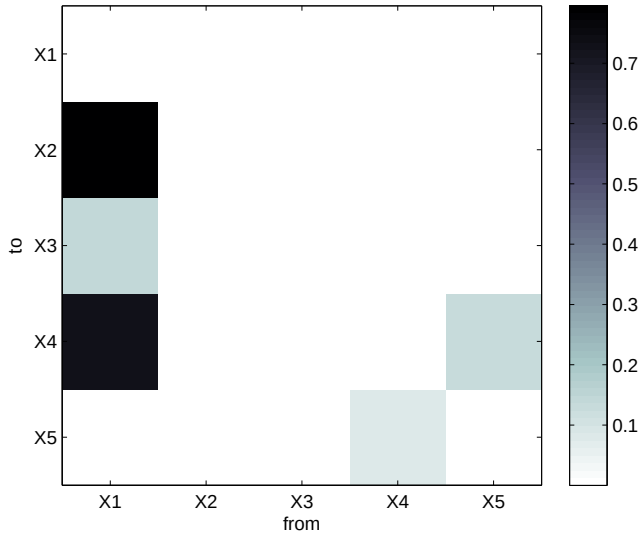


Figure 6.3: Spectral transfer entropy of the simulated data

As shown in Fig. 6.4, while a line with an arrow indicates a unidirectional causality from one variable to the other, the one with double headed arrow suggests a bidirectional causality relationship. From both figures we can see clearly that variable X_1 is the root cause of the oscillation, as validated in the process model shown in Eq. (6.21).

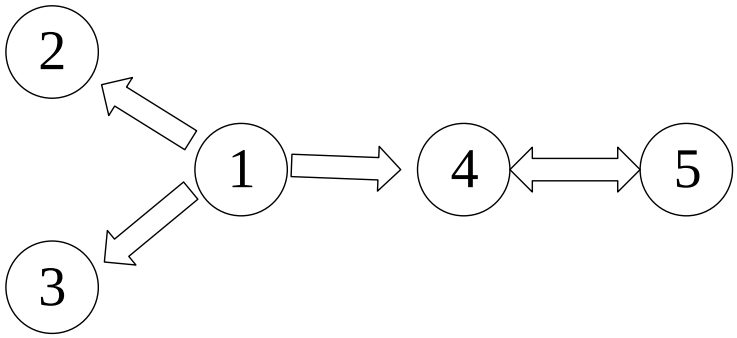


Figure 6.4: Process topology of the simulated data based on spectral transfer entropy

6.3.2 Industrial Data

An industrial data set provided by the Advanced Controls Technology group of Eastman Chemical Company [149, 90]. The process schematic of the plant is shown in Fig. 6.5:

In Fig. 6.5, AC, FC, LC, PC,and TC stand for controlled composition, flow, level, pressure and temperature tags; FI, LI, PI, TI and SI stand for the indicators of flow, level, pressure, temperature and rotor speed. As shown in Fig. 6.5, there are two decanters, three distillation columns and numerous recycle streams in the process. A common disturbance with an oscillation period of about 2 hours (about 320 samples/cycle) had been identified, namely: valve stiction in the actuator of control loop LC2. The feasibility and effectiveness of the spectral transfer entropy method will be demonstrated.

The first 28 h data sampled at every 20s are selected to perform oscillation diagnosis, including 14 controlled process variables (pv's). The normalized

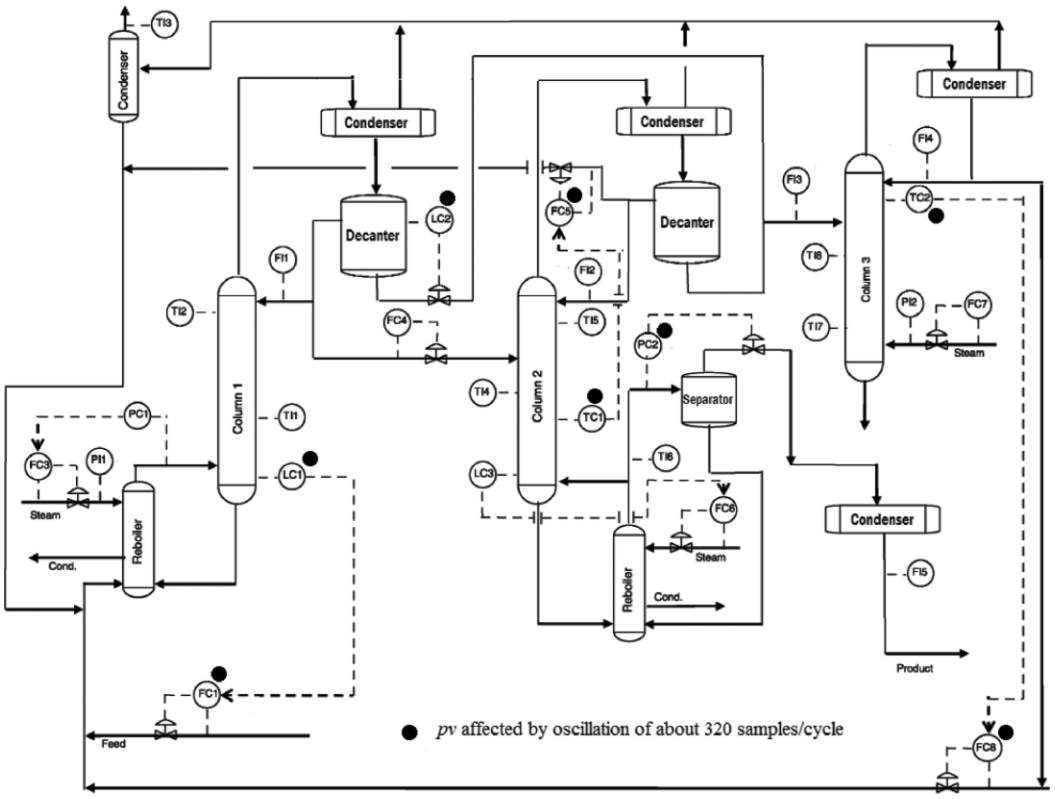


Figure 6.5: Process schematic. The oscillation variables are marked by circle symbols

time trends and power spectra of the data are shown in Fig. 6.6: several variables share a common oscillation at a frequency about 0.003 cycles/sample. Instead of drawing the power spectra for all variables, the spectral envelope method successfully finds a common frequency at 0.0031 cycles/sample (approximately 322 samples/cycle) shown in Fig. 6.7, and related variables with values of $2\left|\hat{\beta}_j(\omega)\right|^2/\sigma_j(\omega)$ larger than $\chi^2_2(0.001) = 13.82$ are listed in Table 6.2:

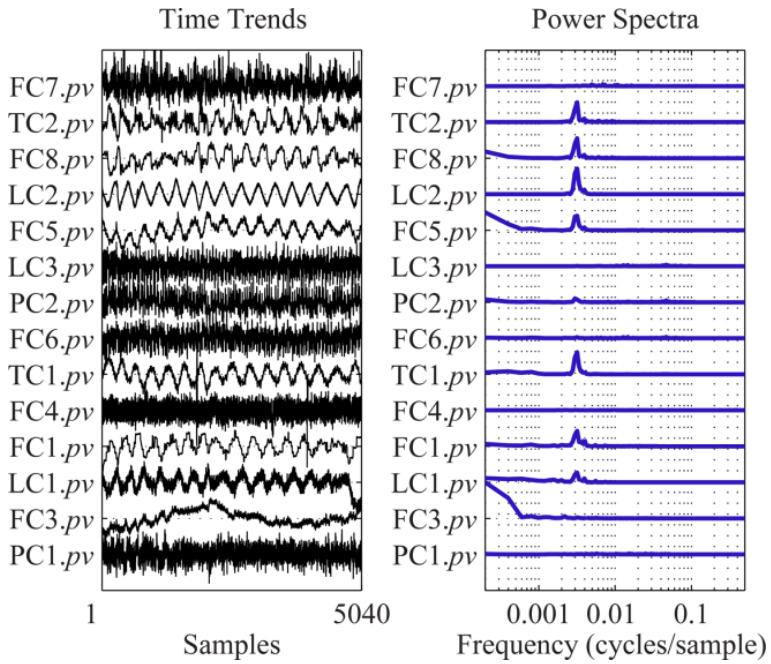


Figure 6.6: Time trends and power spectra of measurements of process variables (pv's)

Combing Fig. 6.6 with Table 6.2, we can see that the eight variables listed are stationary, and their oscillation contribution indices (OCI) calculated from Eq. (6.5) are listed in descending order. The variable LC2.PV has the

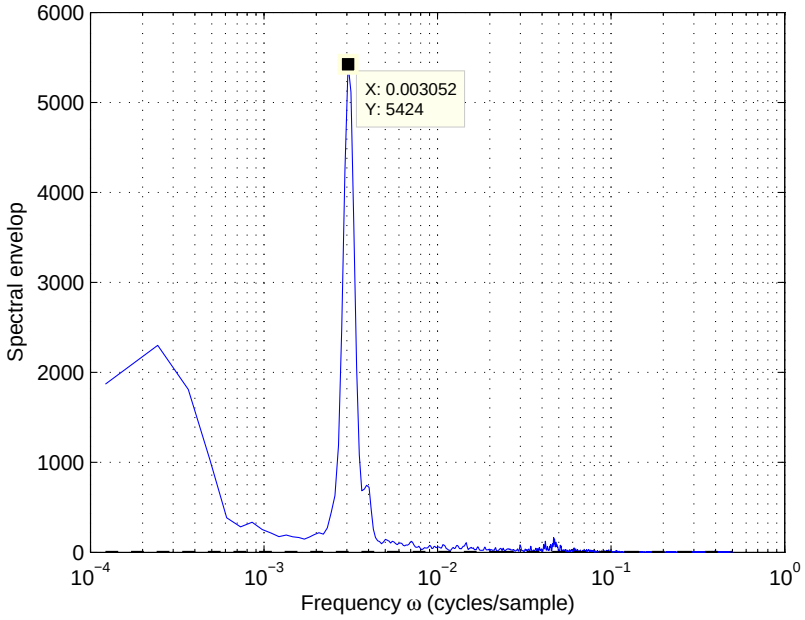


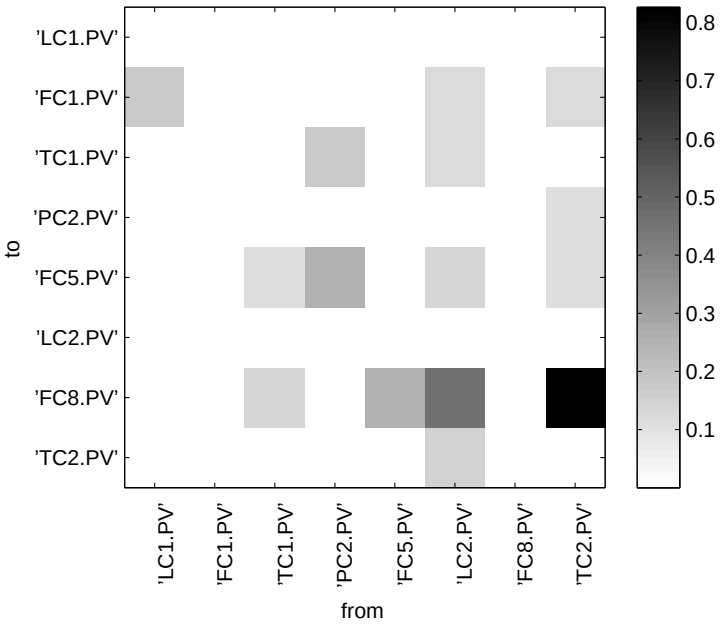
Figure 6.7: Spectral envelope of the Eastman Chemical process data

Table 6.2: Summary of test statistics and oscillation contribution index

Tag no.	Test statistic	OCI	Tag no.	Test statistic	OCI
LC2.PV	706.1	1.62	FC1.PV	941.3	0.36
TC1.PV	1067.4	1.42	LC1.PV	680.9	0.83
FC5.PV	543.2	1.03	FC8.PV	277.8	0.31
PC2.PV	1370.4	0.14	TC2.PV	261.5	0.64

largest value of OCI, which indicates that it may be the source of oscillation.

If one assumes those eight stationary variables are Gaussian, at $f = 0.0031$ cycles/sample, the resulting spectral transfer entropy matrix and process topology diagram are shown in Figs. 6.8 and 6.9, respectively:



E

Figure 6.8: Spectral transfer entropy of the simulated data

A causal map representing the interconnected oscillation propagation pathways is shown in Fig. 6.9. We can see that LC1.pv and LC2.pv do not receive any causal effects from other variables. While LC2.pv reaches all other variables directly or indirectly except LC1.pv, LC1.pv only affects FC1.pv. Thus, we may draw a conclusion that LC2.pv is more likely to be the root cause candidate. Fig. 6.4 also shows that oscillations in loop LC2 propagate

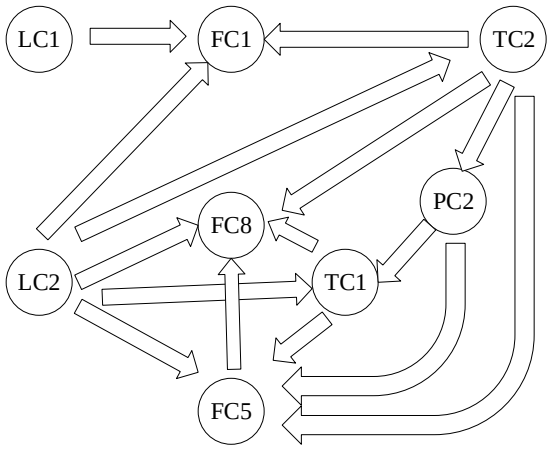


Figure 6.9: Process topology of the Eastman Chemical process based on spectral transfer entropy

to loops FC1, TC2, FC8, TC1, and FC5. By combining this information with the process flow sheet shown in Fig. 6.5, one can conclude that oscillations of loop LC2 propagate from the left-hand decanter to columns 1, 2 and 3, which is consistent with material flow pathways in the physical process. Moreover, the causality between LC1.pv and FC1.pv also matches the cascade control strategy for the liquid level of column 1.

The oscillation period changes vs time can be revealed by the wavelet power spectrum of LC2.PV shown in Fig. 6.10. The period of oscillation gradually reaches 320 samples/cycle near the 1000th sample, and remains stable afterwards.

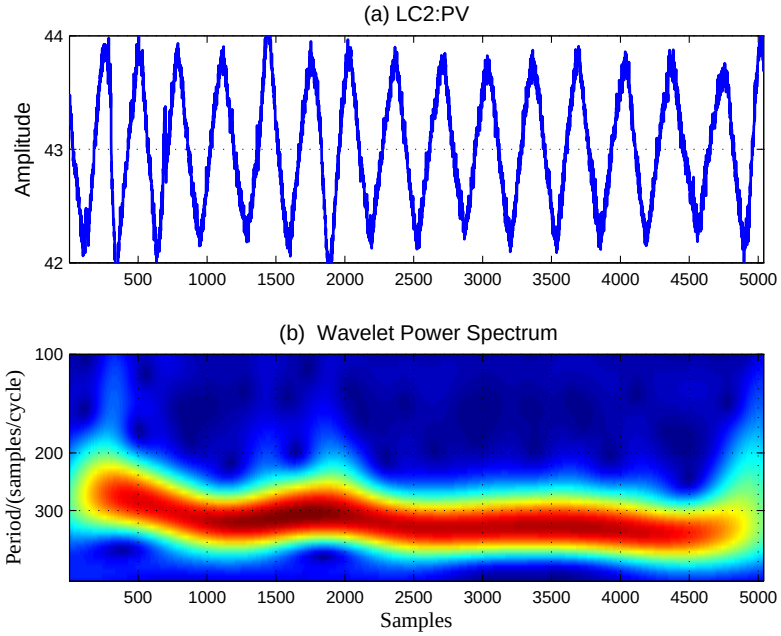


Figure 6.10: Wavelet power spectrum of variable LC2.PV

6.3.3 Discussion

From the results of simulated and industrial cases, we can see that for oscillation detection the spectral envelope method provides a convenient and intuitive way to find the common frequency corresponding to abnormality and related variables. For the root cause diagnosis step, both the oscillation contribution index (OCI) and the spectral transfer entropy give satisfactory results. While the OCI method enjoys the advantage of a low computational cost, it may be hard to find a physical explanation for the diagnosis result and the assumption that root cause variables have relatively larger power at the specific frequency is not validated. Although the spectral transfer entropy provides an intuitive process topology to visualize and explain the causal-effect relationship in the frequency domain, it is based on linear AR model identification and the assumption that the variables are Gaussian is often not satisfied. Thus, the result of oscillation diagnosis in an industrial case study needs to be confirmed through field tests, and the wavelet power spectrum can be used as a promising tool to analyze the period change over time for root cause candidates.

6.4 Summary

As a common type of process disturbance, plant-wide oscillations propagate between units and negatively affect the process. Thus, it is necessary to detect those oscillations and diagnose the root cause. In this chapter, a novel information transfer method is proposed which combines the spectral envelope

with the spectral transfer entropy calculations. In simulated and industrial case studies, for root cause diagnosis, the new method is compared with the oscillation contribution index (OCI). It enjoys the advantage of extracting the process topology and oscillation propagation pathways at a specific frequency range, which is especially useful when the plant data is contaminated with noise at unwanted frequencies.

Chapter 7

Summary and future recommendations

7.1 Summary of contributions

Modern chemical plants are heavily instrumented with IP-enabled intelligent devices. The raw data collected from process industries usually have low quality and have to be cleaned before the model building and knowledge discovery step. This work mainly focuses reviewing (Chapter 2) and improving procedures and tools for data cleaning (Chapters 3$\sim$ 5). Chapter 6 studies a practical problem commonly faced in the process industry — the plant-wide oscillation detection and root cause diagnosis, which facilitates the understanding of data cleaning and knowledge discovery in the process industries.

After reviewing data cleaning methods from both the traditional applied statistics and the state-of-art machine learning discipline in Chapter 2, two suggestions are proposed on improving current data cleaning methodology and discussed: one is incorporating the model performance evaluation and the other is exploring on-line implementation of data cleaning methods. As a premise, in Chapter 3, the negative impact of outlier and noise on the dynamic model identification, and time delays on partial least squares (PLS) model prediction are studied. The simulated and industrial case studies further validate

the necessity to incorporate data cleaning techniques in a knowledge discovery process.

In Chapter 4, an integrated data cleaning scheme is proposed which consists of data cleaning, parameter tuning and model performance estimation. As validated by an industrial case study, the scheme enjoys more versatility than other methods such as RPLS in treating contaminated data sets: various data cleaning techniques can be plugged in the scheme and their parameters will automatically tuned to guarantee the model performance and the quality of data at the same time. Moreover, the new methodology can also circumvent the over-cleaning problem.

In Chapter 5, a time series Kalman filter (TSKF) for outlier detection has been proposed that is suitable for both univariate and multivariate outlier detection in dynamic data sets. Both on-line and off-line versions and related parameter tuning are provided. The TSKF method incorporates the Burg-type time series model fitting algorithm which solves the problem of ill-conditioned auto-covariance matrices. The simulation and plant data testing results demonstrate a superior performance of the TSKF method in additive outlier detection in both an ARMA (1, 1) and a VARMA (1, 1) processes. For innovational outliers, the TSKF only outperforms the Hampel identifier in the univariate case, yet obtains similar results with PCA and DPCA probably because of interactions between IOs and system dynamics.

In Chapter 6, a novel information transfer method which combines spectral envelope with entropy calculations is proposed to detect and diagnose

plant-wide oscillations. Both simulated and industrial case studies demonstrate that the new method can extract process topology and oscillation propagation pathways at a specific frequency range, which is valuable considering that plant data are usually contaminated with noise at unwanted frequencies.

7.2 Recommendations for future work

The research should continue in the following directions:

- Incorporating the data cleaning procedure into data analysis tools to robustify the tool itself. For example, the EM-PCA [273] can handle contaminated data sets with both missing values and outliers, and by incorporating the EM procedure, it can clean the data set and maximize the PCA model performance at the same time. Because machine learning algorithms developed recently excel in data analysis and information extraction, it is important to find more ways to robustify those algorithms.

- In the integrated data cleaning scheme, only univariate data cleaning methods such as the 3σ rule and the Hampel identifier have been applied. In addition, the parameter updating algorithm used to find the optimal model performance is not very effective. Thus, further research should focus on improving the efficiency of parameter updating algorithm and applying multivariate data cleaning methods.

- Integrating the data cleaning procedures with design and tuning of con-

trollers to improve their on-line performance in the presence of signals containing missing values, outliers and noise.

- For the time series Kalman filter (TSKF), improvements should be made so that it will obtain a better performance in non-stationary processes.

- As discussed in the introduction, data collected from the process industries often exhibits a multilevel structure. Although several multi-block and hierarchical PCA and PLS algorithms [316, 239, 83] have been developed to perform data analysis on multilevel data sets, there is a lack of guidance on cleaning data in such a situation. Thus, it would be useful if researchers can develop specific cleaning tools or incorporate them into current algorithms to improve the results for multilevel data sets.

Appendices

Appendix A

Nomenclature

Table A.1: Nomenclature

α	Significance level
β	Mis-identification rate(Type I error)
γ	Normal data estimation rate
Δ	Threshold for outlier identification
ϵ_t	white noise in the ARMA model
$\boldsymbol{\epsilon_t}$	white noise in the VARMA model
θ	The autoregressive model coefficients
Θ	State transition matrix
κ	Normal data estimation rate
μ	Sample mean
σ^2	Sample variance
χ	Outlier detection rate
ϕ	The autoregressive model coefficients
$\mathbf{\Phi}$	Multivariate (vector) autoregressive model coefficient matrix
$\mathbf{\Omega}$	Multivariate (vector) autoregressive model coefficient matrix
AO	Additive outlier

Table A.1: Nomenclature(continued)

$AR(p)$	Autoregressive model with order p
AIC	Akaike information criterion
Amp	Outlier size
$ARMA(p,q)$	Autoregressive moving average model with order p,q
$ARIMA$	Autoregressive integrated moving average model
BIC	Bayesian information criterion
$i.i.d$	Identically and independent distributed
IO	Innovational outlier
LS	Permanent level shift
m	Variable number in a given data set
med	Sample median
MAD	Sample median absolute deviation
$MVAR$	Multivariate (vector) autoregressive model with order p
N	Observation number in a given sample data set
n_p	Number of AR model parameters
rep	Repetition time
TS	Transient level change
$VARMA(p,q)$	Vector autoregressive moving average model with order p, q

Appendix B

The on-line filter-cleaner procedure

Given a univariate process data sequence $\{x_t\}_{t=1}^{N}$ at time t, the filter-cleaner detects outliers on-line following steps below [186]:

1. Choose a data set $\{x_t\}_{t-M+1:t}^{M}$ with window size M.

2. Selection of autoregressive model order r.

3. Estimation of AR(r) model coefficient ϕ based on the data set $\{x_t\}_{t-M+1:t}^{M}$.

   (a) Estimate the mean μ and variance c_0 of $\{x_t\}_{t-M+1:t}^{M}$ based on Hubers M-estimator[137].

   (b) Form new multivariate data sets $\left\{X_i^k = (x_i, x_{i-k})\right\}_{i=t-M+k+1}^{M}$ $(k = 1, 2, ..., r)$. Obtain a robust estimation of the covariance matrix $\begin{bmatrix} .c_{11}^k & c_{12}^k \\ c_{21}^k & c_{22}^k \end{bmatrix}$ of the kth multivariate dataset $\left\{X_i^k = (x_i, x_{i-k})\right\}_{i=t-M+k+1}^{M}$ by the reweighted MCD method[256, 255, 258]. The kth autocorrelation coefficient $\omega_k = c_{12}^k / \sqrt{c_{11}^k c_{22}^k}$.

   (c) Estimation of AR(r) model coefficient ϕ by solving Yule-Walker equations:

$$\omega_j = \phi_1 \omega_{j-1} + \phi_2 \omega_{j-2} + ... + \phi_r \omega_{j-r}, j = 1, ..., r \qquad \text{(B.1)}$$

reformat Equation (B.1):

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_r \end{bmatrix}, \omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_r \end{bmatrix}, P = \begin{bmatrix} 1 & \omega_1 & \cdots & \omega_r \\ \omega_1 & 1 & \cdots & \omega_{r-1} \\ \vdots & \vdots & \vdots & \vdots \\ \omega_r & \omega_{r-1} & \cdots & 1 \end{bmatrix} \quad \text{(B.2)}$$

so that $\phi = P^{-1}\omega$, and the process model can be expressed as:

$$z_t = \frac{\mu}{1 - \phi_1 - \phi_2 - ... - \phi_r} + \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_r z_{t-r} + \varepsilon_t. \quad \text{(B.3)}$$

4. Filter and clean the current data point x_t.

   (a) Reformat the process model in the state-space form:

$$Z_t = \Phi Z_{t-1} + U_t \quad \text{(B.4)}$$

where

$$Z_t^T = [z_t, z_{t-1}, ..., z_{t-r+1}] \quad \text{(B.5)}$$

$$U_t^T = [\varepsilon_t, 0, ..., 0] \quad \text{(B.6)}$$

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{r-1} & \phi_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & & 0 \\ \vdots & 0 & \cdots & \vdots & \vdots \\ \vdots & & \cdots & \vdots & \vdots \\ 0 & & \cdots & 1 & 0 \end{bmatrix} \quad \text{(B.7)}$$

(b) The filter-cleaner computes robust estimates of the vector X_t based on:

$$\hat{Z}_t = \Phi \hat{Z}_{t-1} + \tilde{m}_t s_t \Psi \left(\frac{x_t - \hat{x}_t^{t-1}}{s_t} \right) \tag{B.8}$$

where $\tilde{m}_t = m_t / s_t^2$, and $\tilde{m}_t$ is the first column of M_t:

$$M_{t+1} = \Phi P_t \Phi^T + Q \tag{B.9}$$

$$P_t = M_t - \pi \left(\frac{x_t - \hat{x}_t^{t-1}}{s_t} \right) \tag{B.10}$$

where Q is a matrix with all zero entries except $Q_{11} = \sigma_\varepsilon^2$. $s_t^2 = m_{11,t}.\hat{x}_t^{t-1}$ denotes a robust one-step ahead prediction of x_t and $\hat{x}_t^{t-1} = \left(\Phi \hat{Z}_{t-1} \right)_1$. The psi-function,$\Psi$ and weight function,π are chosen to be:

$$\Psi(\tau) = \begin{cases} \tau, & |\tau| < 3 \\ 0, & |\tau| \geq 3 \end{cases} \tag{B.11}$$

$$\pi(\tau) = \frac{\Psi(\tau)}{\tau} \tag{B.12}$$

(c) Finally, the cleaned data at time t is given by:

$$\hat{z}_t = \left(\hat{Z}_t \right)_1 \tag{B.13}$$

.

Appendix C

Proof for the Fourier transform of Gaussian variables

Assume variable X follows a Gaussian distribution $X \sim f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, its Fourier transform is given by:

$$\begin{aligned} &F_x\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\right](\omega) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} e^{-i\omega x} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \left[\cos(\omega x) - i\sin(\omega x)\right] dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \cos(\omega x)\, dx - i\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \sin(\omega x)\, dx\right] \end{aligned} \tag{C.1}$$

Since the second integrand is odd, so integration over a symmetrical range gives zero. Calculating the first integral we have:

$$F_x[X](\omega) = 2\frac{1}{\sigma\sqrt{2\pi}}\sqrt{2\pi}\sigma\frac{1}{2}e^{-\sigma^2\omega^2/2} = \exp\left(-\frac{\omega^2\sigma^2}{2}\right) \tag{C.2}$$

Thus, we can see that the Fourier transform of a Gaussian is also Gaussian, and the result can be extended to a multivariate case.

Bibliography

[1] Bovas Abraham and George E. P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–236, 1979.

[2] Bovas Abraham and Alice Chuang. Outlier detection and time series modeling. *Technometrics*, 31(2):241–248, 1989.

[3] Z. H. Abuelzeet, V. M. Becerra, and P. D. Roberts. Combined bias and outlier identification in dynamic data reconciliation. *Comput Chem Eng*, 26(6):921 – 935, 2002.

[4] Luís Aguiar-Conraria and Maria Joana Soares. The continuous wavelet transform: A primer. Technical report, Economics Department, University of Minho, Braga, Portugal,, 2011.

[5] Salim Ahmed. *Parameter and delay estimation of continuous-time models from uniformly and non-uniformly sampled data*. PhD thesis, University of Alberta, Edmonton, Canada,, 2006.

[6] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[7] João S. Albuquerque and Lorenz T. Biegler. Data reconciliation and gross-error detection for dynamic systems. *AIChE J.*, 42(10):2841–2856, 1996.

[8] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[9] Paul D. Allison. Handling missing data by maximum likelihood. In *SAS Global Forum Satistics and Data Analysis*, pages 1–21, 2012.

[10] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais, and S.J. Formosinho. Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometr Intell Lab Syst*, 87(2):208 – 217, 2007.

[11] Jaafar AlMutawa. Identification of errors-in-variables state space models with observation outliers based on minimum covariance determinant. *J Process Control*, 19(5):879 – 887, 2009.

[12] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat*, 46(3):175–185, 1992.

[13] Mina Aminghafari, Nathalie Cheze, and Jean-Michel Poggi. Multivariate denoising using wavelets and principal component analysis. *Comput Stat Data Anal*, 50(9):2381 – 2398, 2006. Statistical signal extraction and filtering Statistical signal extraction and filtering.

[14] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial

basis function approach to nonlinear granger causality of time series. *Phys. Rev. E*, 70:056221, Nov 2004.

[15] Brian D.O. Anderson and John B. Moore. *Optimal Filtering*. NJ: Prentice Hall, 1979.

[16] Robert Anderson. *Modern methods for roust regression*. Quantitative Applications in the Social Sciences. New York: SAGE Publications, Inc., 2008.

[17] Francisco Arteaga and Alberto Ferrer. Dealing with missing data in mspc: several methods, different interpretations, some examples. *J. Chemom.*, 16(8-10):408–418, 2002.

[18] Francisco Arteaga and Alberto Ferrer. Framework for regression-based missing data imputation methods in on-line mspc. *J. Chemom.*, 19(8):439–447, 2005.

[19] Karl Johan Åström and Björn Wittenmark. *Adaptive Control*. Addison-Wesley, Reading, Massachusetts, January 1995.

[20] Luiz A. Baccalá and Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biol.Cybern.*, 84(6):463–474, 2001.

[21] Bhavik R. Bakshi. Multiscale pca with application to multivariate statistical process monitoring. *AIChE J.*, 44(7):1596–1610, 1998.

[22] Amanda N. Baraldi and Craig K. Enders. An introduction to modern missing data analyses. *J Sch Psychol*, 48(1):5 – 37, 2010.

[23] P. Baraldi, F. Di Maio, D. Genini, and E. Zio. Reconstruction of missing data in multidimensional time series by fuzzy similarity. *Appl Soft Comput*, 26:1–9, 2014.

[24] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.*, 103:238701, Dec 2009.

[25] Lionel Barnett and Anil K. Seth. The MVGC multivariate granger causality toolbox: A new approach to granger-causal inference. *J. Neurosci. Meth.*, 223(0):50 – 68, 2014.

[26] Vic Barnett and Toby Lewis. *Outliers in statistical data*. Wiley Series in Probability and Mathematical Satistics. Chichester:Wiley,, 2nd edition, 1984.

[27] M. Bauer, J.W. Cox, M.H. Caveness, James J. Downs, and N.F. Thornhill. Finding the direction of disturbance propagation in a chemical process using transfer entropy. *IEEE Trans. Control Syst. Technol.*, 15(1):12–21, Jan 2007.

[28] Margret Bauer, John W. Cox, Michelle H. Caveness, James J. Downs, and Nina F. Thornhill. Nearest neighbors methods for root cause analysis of plantwide disturbances. *Ind.Eng. Chem.Res.*, 46(18):5977–5984, 2007.

[29] Margret Bauer and Nina F. Thornhill. A practical method for identifying the propagation path of plant-wide disturbances. *J. Process Contr.*, 18(7C8):707 – 719, 2008.

[30] Vinay A. Bavdekar, Anjali P. Deshpande, and Sachin C. Patwardhan. Identification of process and measurement noise covariance for state and parameter estimation using extended kalman filter. *J Process Control*, 21(4):585 – 601, 2011.

[31] Claudia Becker. The size of the largest nonidentifiable outlier as a performance criterion for multivariate outlier identification: the case of high-dimensional data. In Jelke G. Bethlehem and Peter G.M. van der Heijden, editors, *COMPSTAT*, pages 211–216. Heidelberg: Physica-Verlag,, 2000.

[32] J. Benesty, Yiteng Huang, and Jingdong Chen. Time delay estimation via minimum entropy. *IEEE Signal Process Lett*, 14(3):157–160, March 2007.

[33] A. M. Bianco, M. Garca Ben, E. J. Martnez, and V. J. Yohai. Outlier detection in regression models with arima errors using robust estimates. *J Forecasting*, 20(8):565–579, 2001.

[34] C. M. Bishop. Novelty detection and neural network validation. *Vision, Image and Signal Processing, Proc IEE*, 141(4):217–222, Aug 1994.

[35] Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. New York: Springer-Verlag,, 2006.

[36] Christopher M. Bishop, Markus Svensén, and Christopher K.I. Williams. Gtm: The generative topographic mapping. *Neural Comput.*, 10(1):215–234, 1998.

[37] Svante Björklund. A survey and comparison of time-delay estimation methods in linear systems. Technical report, Linkopings University, 2003.

[38] Terry Blevins, Mark Nixon, and Marty Zielinski. Using wireless measurements in control applications. Technical report, Emerson Process Management, 2013.

[39] C.a. Bode, B.S. Ko, and T.F. Edgar. Run-to-run control and performance monitoring of overlay in semiconductor manufacturing. *Control Eng Pract*, 12(7):893–900, July 2004.

[40] Andrey Bogomolov. Multivariate process trajectories: capture, resolution and analysis. *Chemometr Intell Lab Syst*, 108(1):49 – 63, 2011.

[41] Richard J. Bolton and David J. Hand. Unsupervised profiling methods for fraud detection. In *Proc Credit Scoring and Credit Control VII*, pages 5–7, 2001.

[42] Fani Boukouvala, Fernando J. Muzzio, and Marianthi G. Ierapetritou. Predictive modeling of pharmaceutical processes with missing and noisy data. *AIChE J.*, 56(11):2860–2872, 2010.

[43] G. E. P Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forcasting and control*. New York:Wiley, 4th edition, 2013.

[44] P. S. Bradley, Usama M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: formulations and challenges. *INFORMS J Comput*, 11(3):217–238, 1999.

[45] Leo Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

[46] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000.

[47] Robert Grover Brown and Patrick Y.C. Hwang. *Introduction to random signals and applied Kalman filtering*. New Jersey: John Wiley & Sons, Ltd.,, 4th edition, 2012.

[48] Simon Byers and Adrian E. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *J. Am. Stat. Assoc.*, 93(442):577–584, 1998.

[49] Qiao Cai, Haibo He, and Hong Man. Spatial outlier detection based on iterative self-organizing learning model. *Neurocomputing*, 117(0):161 – 172, 2013.

[50] José Camacho. Missing-data theory in the context of exploratory data analysis. *Chemometr Intell Lab Syst*, 103(1):8 – 18, 2010.

[51] José Camacho. Visualizing big data with compressed score plots: approach and research challenges. *Chemometr Intell Lab Syst*, 135(0):110 – 125, 2014.

[52] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[53] Kathryn Chaloner and Rollin Byant. A bayesian approach to outlier detection and residual analysis. *Biometrika*, 75(4):651–659, 1988.

[54] Ih Chang, George C. Tiao, and Chung Chen. Estimation of time series parameters in the presence of outliers. *Technometrics*, 30(2):193–204, 1988.

[55] Peter Cheeseman, Matthew Self, Jim Kelly, Will Taylor, Don Freeman, and John Stutz. Bayesian classification. In *Proceedings of American Association of Artificial Intelligence (AAAI)*, pages 607–611. San Mateo: Morgan kaufmann, 1988.

[56] Chung Chen and Lon-Mu Liu. Joint estimation of model parameters and outlier effects in time series. *J. Am. Stat. Assoc.*, 88(421):284–297, 1993.

[57] J. Chen, A. Bandoni, and J.A. Romagnoli. Robust statistical process monitoring. *Comput Chem Eng*, 20, Supplement 1(0):497– 502, 1996. European Symposium on Computer Aided Process Engineering-6.

[58] J. Chen, A. Bandoni, and J.A. Romagnoli. Outlier detection in process plant data. *Comput Chem Eng*, 22(4C5):641 – 646, 1998.

[59] J. Chen and J.A. Romagnoli. A strategy for simultaneous dynamic data reconciliation and outlier detection. *Comput Chem Eng*, 22(4C5):559 – 562, 1998.

[60] Tao Chen, Julian Morris, and Elaine Martin. Dynamic data rectification using particle filters. *Comput Chem Eng*, 32(3):451 – 462, 2008.

[61] Wen-shiang Chen. *Bayesian estimation by sequential Monte Carlo sampling for nonlinear dynamic systems*. PhD thesis, The Ohio State University, 2004.

[62] Zhe Chen. Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182:1–69, 2003.

[63] Leo H. Chiang, Randy J. Pell, and Mary Beth Seasholtz. Exploring process data with the use of robust outlier detection algorithms. *J Process Control*, 13(5):437 – 449, 2003.

[64] Leo H. Chiang, Evan L. Russell, and Richard D. Braatz. *Fault detection and diagnosis in industrial systems*. London:Springer-Verlag, 2001.

[65] Ji-Hoon Cho, Jong-Min Lee, Sang Wook Choi, Dongkwon Lee, and In-Beum Lee. Fault identification for process monitoring using kernel principal component analysis. *Chem Eng Sci*, 60(1):279 – 288, 2005.

[66] Sang Wook Choi, Changkyu Lee, Jong-Min Lee, Jin Hyun Park, and In-Beum Lee. Fault detection and identification of nonlinear processes based on kernel {PCA}. *Chemometr Intell Lab Syst*, 75(1):55 – 67, 2005.

[67] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometr Intell Lab Syst*, 78:103 – 112, 2005.

[68] Anders Christoffersson. *The one component model with incomplete data*. PhD thesis, Uppsala University, 1970.

[69] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, 1994.

[70] Corinna Cortes and Vladimir Vapnik. Support vector networks. *Mach Learn*, 20(3):273–297, 1995.

[71] C. Croux, P. Filzmoser, and M.R. Oliveira. Algorithms for projection-cpursuit robust principal component analysis. *Chemometr Intell Lab Syst*, 87(2):218 – 225, 2007.

[72] Christophe Croux, Peter J. Rousseeuw, and Ola Hössjer. Generalized s-estimators. *J. Am. Stat. Assoc.*, 89(428):1271–1281, 1994.

[73] Domenico Cucina, Antonietta di Salvatore, and Mattheos K. Protopapas. Outliers detection in multivariate time series using genetic algorithms. *Chemometr Intell Lab Syst*, 132(0):103 – 110, 2014.

[74] Wentong Cui and Xuefeng Yan. Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. *Chemometr Intell Lab Syst*, 98(2):130 – 135, 2009.

[75] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak. Robust statistics in data analysis a review: Basic concepts. *Chemometr Intell Lab Syst*, 85(2):203 – 219, 2007.

[76] Laurie Davies and Ursula Gather. The identification of multiple outliers. *J. Am. Stat. Assoc.*, 88(423):782–792, 1993.

[77] Jim Davis, Thomas F. Edgar, Jim Porter, John Bernaden, and Michael S. Sarli. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Comput Chem Eng*, 47:145–156, 2012.

[78] Bhupinder S. Dayal and John F. MacGregor. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J Process Control*, 7(3):169 – 179, 1997.

[79] M. J. L. de Hoon, T. H. J. J. van der Hagen, H. Schoonewelle, and H. van Dam. Why yule-walker should not be used for autoregressive modelling. *Annals of Nuclear Energy*, 23(15):1219–1228, 1996.

[80] Sijmen De Jong. Simpls: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*, 18(3):251–263, 1993.

[81] Rodrigo López-Negrete de la Fuente, Salvador Garcí a Muñoz, and Lorenz T. Biegler. An efficient nonlinear programming strategy for pca models with incomplete data sets. *J. Chemom.*, 24(6):301–311, 2010.

[82] C. L. de Ligny, G. H. E. Nieuwdorp, W. K. Brederode, W. E. Hammers, and J. C. van Houwelingen. An application of factor analysis with missing data. *Technometrics*, 23(1):91–95, 1981.

[83] Onno E. de Noord and Eugene H. Theobald. Multilevel component analysis and multilevel pls of chemical process data. *J. Chemom.*, 19(5-7):301–307, 2005.

[84] A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*, 39(1):1–38, 1977.

[85] Jing Deng and Biao Huang. Identification of nonlinear parameter varying systems with missing output data. *AIChE J.*, 58(11):3454–3467, 2012.

[86] Alessandro G. Di Nuovo. Missing data analysis with fuzzy C-means: A study of its application in a psychological scenario. *Expert Syst Appl*, 38:6793 – 6797, 2011.

[87] Terry E Dielman. Least absolute value regression: recent contributions. *J Stat Comput Simul*, 75(4):263–286, 2005.

[88] Scott C. Douglas. *Digital Signal Processing Handbook*, chapter Introduction to Adaptive Filters. Boca Raton: CRC Press LLC, 1999.

[89] Fuat Doymaz, Amid Bakhtazad, Jose A. Romagnoli, and Ahmet Palazoglu. Wavelet-based robust filtering of process data. *Comput Chem Eng*, 25(11C12):1549 – 1559, 2001.

[90] Ping Duan, Tongwen Chen, Sirish L. Shah, and Fan Yang. Methods for root cause diagnosis of plant-wide oscillations. *AIChE J.*, 60(6):2019–2034, 2014.

[91] Ping Duan, Fan Yang, Tongwen Chen, and S.L. Shah. Direct causality detection via the transfer entropy approach. *IEEE Trans. Control Syst. Technol.*, 21(6):2052–2066, Nov 2013.

[92] Emil Eirola, Gauthier Doquire, Michel Verleysen, and Amaury Lendasse. Distance estimation in numerical data sets with missing values. *Inform Sci*, 240(0):115 – 128, 2013.

[93] Emil Eirola, Amaury Lendasse, Vincent Vandewalle, and Christophe Biernacki. Mixture of Gaussians for distance estimation with missing data. *Neurocomputing*, 131(0):32 – 42, 2014.

[94] Anders Eriksson and Anton Van Den Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l 1 norm. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.*, pages 771–778. IEEE, 2010.

[95] Kim H Esbensen, Maths Halstensen, Thorbjørn Tønnesen Lied, Arild Saudland, Jørild Svalestuen, Sunil de Silva, and Bjørn Hope. Acoustic

chemometricsfrom noise to information. *Chemometr Intell Lab Syst*, 44(1C2):61 – 76, 1998.

[96] Hector Joaquin Galicia Escobar. *Advanced monitoring and soft sensor development with application to industrial processes*. PhD thesis, Auburn University, Auburn, Alabama, USA,, May 2012.

[97] Martin Ester, Hans peter Kriegel, J?rg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[98] C. Faloutsos, F. Korn, A. Labrinidis, Y. Kotidis, A. Kaplunovich, and D. Perković. Quantifiable data mining using principal component analysis. Technical Report Technical Report 97-25, Institute for Systems Research, University of Maryland, College Park, MD, 1997.

[99] J. A. Fernández-Pierna, L. Jin, M. Daszykowski, F. Wahl, and D. L. Massart. A methodology to detect outliers/inliers in prediction with PLS. *Chemometr Intell Lab Syst*, 68(1C2):17 – 28, 2003.

[100] J. A. Fernández-Pierna, F. Wahl, O. E. de Noord, and D. L Massart. Methods for outlier detection in prediction. *Chemometr Intell Lab Syst*, 63(1):27 – 39, 2002.

[101] P. Filzmoser, C. Dehon, and C. Croux. Outlier resistant estimators for canonical correlation analysis. In JelkeG. Bethlehem and PeterG.M. van der Heijden, editors, *COMPSTAT*, pages 301–306. Heidelberg: Physica-Verlag,, 2000.

[102] B.R. Fischer and A. Medvedev. L2 time delay estimation by means of laguerre functions. In *Procedings of American Control Conference*, pages 455–459, 1999.

[103] A.J. Fox. Outliers in time series. *J R Stat Soc Series B Stat Methodol*, 34(3):350–363, 1972.

[104] Philip Hans Franses and Andr Lucas. Outlier detection in cointegration analysis. *J Bus Econ Stat*, 16(4):459–468, 1998.

[105] Michael Frigge, David C. Hoaglin, and Boris Iglewicz. Some implementations of the boxplot. *Am Stat*, 43(1):50–54, 1989.

[106] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans Inf Theory*, 21(1):32–40, Jan 1975.

[107] K. Ruben Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, 1979.

[108] Pedro Galeano, Daniel Peña, and Ruey S. Tsay. Outlier detection in multivariate time series by projection pursuit. *J. Am. Stat. Assoc.*, 101(474):654–669, 2006.

[109] Roberto Kawakami Harrop Galvão, Gledson Emídio José, Heronides Adonias Dantas Filho, Mario Cesar Ugulino Araujo, Edvan Cirino da Silva, Henrique Mohallem Paiva, Teresa Cristina Bezerra Saldanha, and Ênio Sartre Oliveira Nunes de Souza. Optimal wavelet filter construction using x and y data. *Chemometr Intell Lab Syst*, 70(1):1 – 10, 2004.

[110] Zhiqiang Ge. Quality prediction and analysis for large-scale processes based on multi-level principal component modeling strategy. *Control Eng Pract*, 31(0):9 – 23, 2014.

[111] Zhiqiang Ge, Chunjie Yang, and Zhihuan Song. Improved kernel pca-based monitoring approach for nonlinear processes. *Chem Eng Sci*, 64(9):2245 – 2255, 2009.

[112] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Anal Chim Acta*, 185(0):1 – 17, 1986.

[113] John Geweke. Measurement of linear dependence and feedback between multiple time series. *J. Am. Statist. Assoc.*, 77(378):304–313, 1982.

[114] M.P. Gómez-Carracedo, J.M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada. A practical comparison of single and multiple imputa-

tion methods to handle complex missing data in air quality datasets. *Chemometr Intell Lab Syst*, 134(0):23 – 33, 2014.

[115] R.B. Govindan, J. Raethjen, F. Kopper, J.C. Claussen, and G. Deuschl. Estimation of time delay by coherence analysis. *Phys. A*, 350(2C4):277 – 295, 2005.

[116] John W. Graham. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.*, 60:549–576, 2009.

[117] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969.

[118] Bjørn Grung and Rolf Manne. Missing values in principal component analysis. *Chemometr Intell Lab Syst*, 42(1-2):125 – 139, 1998.

[119] Maya R. Gupta and Yihua Chen. *Theory and use of the EM Algorithm*. Foundations and Trends in Signal Processing. Norwell, MA: Now Publishers Inc.,, 2011.

[120] Frank R. Hampel. A general qualitative definition of robustness. *Ann Math Stat*, 42(6):1887–1896, 1971.

[121] Frank R. Hampel. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, 69(346):383–393, 1974.

[122] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. The Morgan Kaufmann Series in Data Management Systems. San Francisco: Morgan kaufmann,, 3rd edition edition, 2006.

[123] Anders Hansson and Ragnar Wallin. Maximum likelihood estimation of Gaussian models with missing data–eight equivalent formulations. *Automatica*, 48(9):1955 – 1962, 2012.

[124] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*, 28(1):100–108, 1979.

[125] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier detection using replicator neural networks. In Yahiko Kambayashi, Werner Winiwarter, and Masatoshi Arikawa, editors, *Data Warehousing and Knowledge Discovery*, volume 2454 of *Lecture Notes in Computer Science*, pages 170–180. Berlin Heidelberg:Springer,, 2002.

[126] Simon Haykin. *Kalman filtering and neural networks*. New York: Wiley,, 2001.

[127] Simon Haykin. *Adaptive filter theory*. Prentince Hall Information and system sciences series. NJ: Prentice Hall,, 5th edition, 2013.

[128] Simon Haykin and Bernard Widrow. *Least-mean-square adaptive filters*. New Jersey: John Wiley & Sons, Ltd.,, 2003.

[129] Simon O. Haykin. *Adaptive filter theory*. NJ: Prentice Hall, 5th edition, 2013.

[130] G. Hinton. Deep belief network. *Scholarpedia*, 4:5947, 2009.

[131] Katerina Hlaváčková-Schindler, Milan Palušand Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.*, 441(1):1 – 46, 2007.

[132] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *AI Rev*, 22(2):85–126, 2004.

[133] Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun Stat Theory Methods*, 6(9):813–827, 1977.

[134] K.A. Hoo, K.J. Tvarlapati, M.J. Piovoso, and R. Hajare. A method of robust multivariate outlier replacement. *Comput Chem Eng*, 26(1):17 – 39, 2002.

[135] Harold Hotelling. The generalization of student's ratio. *Ann Math Stat*, 2(3):360–378, 1931.

[136] Yuzhu Hu, J. Smeyers-Verbeke, and D.L. Massart. Outlier detection in calibration. *Chemometr Intell Lab Syst*, 9(1):31 – 44, 1990.

[137] Peter J Huber. Robust estimation of a location parameter. *Ann Math Stat*, 35(1):73–101, 1964.

[138] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistcs*. New York: Wiley, 2nd edition, 2009.

[139] M. Hubert and K. Vanden Branden. Robust methods for partial least squares regression. *J. Chemom.*, 17(10):537–549, 2003.

[140] Lynette Hunt and Murray Jorgensen. Mixture model clustering for mixed data with missing information. *Comput Stat Data Anal*, 41(3-4):429 – 440, 2003. Recent Developments in Mixture Model.

[141] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw*, 13(4C5):411 – 430, 2000.

[142] S. A. Imtiaz and S. L. Shah. Treatment of missing values in process data analysis. *Can J Chem Eng*, 86(5):838–858, 2008.

[143] A.J. Isaksson, A. Horch, and G.A. Dumont. Event-triggered deadtime estimation from closed-loop data. 4:3280–3285, 2001.

[144] Louis A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann Math Stat*, 43(5):1449–1458, 10 1972.

[145] D. Janakiram, V. Adi Mallikarjuna Reddy, and A.V.U. Phani Kumar. Outlier detection in wireless sensor networks using bayesian belief networks. In *Communication System Software and Middleware, 2006. Comsware 2006. First International Conference on*, pages 1–6, 2006.

[146] Nathalie Japkowicz, Catherine Myers, and Mark Gluck. A novelty detection approach to classification. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, pages 518–523, 1995.

[147] Jin-Tsong Jeng, Chen-Chia Chuang, and Chin-Wang Tao. Hybrid svmr-gpr for modeling of chaotic time series systems with noise and outliers. *Neurocomputing*, 73(10C12):1686 – 1693, 2010. Subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[148] María Jesús Sánchez and Daniel Peña. The identification of multiple outliers in arima models. *Commun Stat Theory Methods*, 32(6):1265–1287, 2003.

[149] Hailei Jiang, M.A.A. Shoukat Choudhury, and Sirish L. Shah. Detection and diagnosis of plant-wide oscillations from industrial data using the spectral envelope method. *J. Process Contr.*, 17(2):143 – 155, 2007.

[150] Hailei Jiang, Rohit Patwardhan, and Sirish L. Shah. Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix. *J. Process Contr.*, 19(8):1347 – 1354, 2009.

[151] Wei Jiang, Zhi-Min Zhang, YongHuan Yun, De-Jian Zhan, Yi-Bao Zheng, Yi-Zeng Liang, Zhen Yu Yang, and Ling Yu. Comparisons of five algorithms for chromatogram alignment. *Chromatographia*, 76(17-18):1067–1078, July 2013.

[152] Julie Josse, Marieke E. Timmerman, and Henk A.L. Kiers. Missing values in multi-level simultaneous component analysis. *Chemometr Intell Lab Syst*, 129(0):21 – 32, 2013. Multiway and Multiset Methods.

[153] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1 – 10, 1991.

[154] Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven soft sensors in the process industry. *Comput Chem Eng*, 33(4):795 – 814, 2009.

[155] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *J Fluids Eng*, 82(1):35–45, 1960.

[156] M.J. Kamiński and K.J. Blinowska. A new method of the description of the information flow in the brain structures. *Biol.Cybern.*, 65(3):203–210, 1991.

[157] Thomas W. Karjala and David M. Himmelblau. Dynamic rectification of data via recurrent neural nets and the extended kalman filter. *AIChE J.*, 42(8):2225–2239, 1996.

[158] Athanassios Kassidas, John F Macgregor, and Paul A Taylor. Synchronization of batch trajectories using dynamic time warping. *AIChE J.*, 44(4):864–875, 1998.

[159] S. M. Kay. *Modern spectral analysis with applications*. NJ: Prentice Hall, 1988.

[160] Shima Khatibisepehr and Biao Huang. Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Ind Eng Chem Res*, 47(22):8713–8723, 2008.

[161] Shima Khatibisepehr and Biao Huang. A bayesian approach to robust process identification with arx models. *AIChE J.*, 59(3):845–859, 2013.

[162] Jae-On Kim and James Curry. The treatment of missing data in multivariate analysis. *Sociol Methods Res*, 6(2):215–240, 1977.

[163] C. Knapp and G.Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Trans Acoust*, 24(4):320–327, Aug 1976.

[164] Edwin M Knorr and Raymond T Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer, 1998.

[165] Teuvo Kohonen. *Self-organizing maps*. Springer Series in Information Sciences. Heidelberg: Physica-Verlag, 3rd edition, 1999.

[166] A Kolmogoroff. Interpolation und extrapolation von stationaren zufalligen folgen. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 5(1):3–14, 1941.

[167] Theodora Kourti. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. *Annu Rev Control*, 27(2):131–139, January 2003.

[168] Theodora Kourti and John F. MacGregor. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometr Intell Lab Syst*, 28(1):3 – 21, 1995.

[169] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1649–1652, New York, NY, USA,, 2009. ACM.

[170] Hans-Peter Kriegel, Peer Kröger, Erich Shubert, and Arthur Zimek. Interpreting and unifying outlier scores. In *Proceedings of 11th SIAM International Conference on Data Mining*, 2011.

[171] Hans-Peter Kriegel, Matthias S hubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 444–452, New York, NY, USA, 2008. ACM.

[172] Wenfu Ku, Robert H. Storer, and Christos Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometr Intell Lab Syst*, 30(1):179 – 196, 1995. InCINC '94 Selected papers from the First International Chemometrics Internet Conference.

[173] Doh-Soon Kwak and Kwang-Jae Kim. A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Syst Appl*, 39(3):2590 – 2596, 2012.

[174] Kamakshi Lakshminarayan, StevenA. Harp, and Tariq Samad. Imputation of missing data in industrial databases. *Appl Intell*, 11(3):259–275, 1999.

[175] LonginJan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Computer Science*, pages 61–75. Berlin Heidelberg:Springer, 2007.

[176] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 157–166, New York, NY, USA, 2005. ACM.

[177] Jaeshin Lee, Bokyoung Kang, and Suk-Ho Kang. Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *J Process Control*, 21(7):1011 – 1021, 2011.

[178] Joon Lee, Shamim Nemati, Ikaro Silva, Bradley Edwards, James Butler, and Atul Malhotra. Transfer entropy estimation and directional coupling change detection in biomedical time series. *Biomed. Eng. Online*, 11(1):19, 2012.

[179] M.J. Leibman, T.F. Edgar, and L.S. Lasdon. Efficient data reconciliation and estimation for dynamic processes using nonlinear programming

techniques. *Comput Chem Eng*, 16(10C11):963 – 986, 1992. An International Journal of Computer Applications in Chemical Engineering.

[180] Weihua Li, Abhishek Bhargava, and Sirish L. Shah. Adaptive process monitoring via multichannel eiv lattice filters. *AIChE J.*, 48(4):786–799, 2002.

[181] M.J. Liebman. *Reconciliation of process measurements using statistical and nonlinear programming techniques*. PhD thesis, University of Texas at Austin, Austin, TX, USA,, 1991.

[182] Bao Lin, Bodil Recke, Jørgen K.H. Knudsen, and Sten Bay Jørgensen. A systematic approach for soft sensor development. *Comput Chem Eng*, 31(5C6):419 – 425, 2007. ESCAPE-15 Selected Papers from the 15th European Symposium on Computer Aided Process Engineering held in Barcelona, Spain, May 29-June 1, 2005.

[183] Michael Lindner, Raul Vicente, Viola Priesemann, and Michael Wibral. Trentool: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neuroscience*, 12(1):119, 2011.

[184] Roderick J. A. Little. Missing-data adjustments in large surveys. *J Bus Econ Stat*, 6(3):pp. 287–296, 1988.

[185] Roderick J. A. Little and Ronald B. Rubin. *Statistical analysis with missing data*. New York:Wiley, 2nd edition edition, September 2002.

[186] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Comput Chem Eng*, 28(9):1635 – 1647, 2004.

[187] Yi Liu and Junghui Chen. Correntropy kernel learning for nonlinear system identification with outliers. *Ind Eng Chem Res*, 53(13):5248–5260, 2014.

[188] Greta M. Ljung. On outlier detection in time series. *J R Stat Soc Series B Stat Methodol*, 55(2):559–567, 1993.

[189] Vitor V. Lopes and José C. Menezes. Inferential sensor design in the presence of missing data: a case study. *Chemometr Intell Lab Syst*, 78(1-2):1 – 10, 2005.

[190] Ricardo A. Losada. *Digitial fIlters with MATLAB*. The MathWorks, Inc, 2008.

[191] Bo Lu, Ivan Castillo, Leo Chiang, and Thomas F. Edgar. Industrial PLS model variable selection using moving window variable importance in projection. *Chemometr Intell Lab Syst*, 135(0):90 – 109, 2014.

[192] Helmut Lütkepohl, Pentti Saikkonen, and Carsten Trenkler. Testing for the cointegrating rank of a var process with level shift at unknown time. *Econometrica*, 72(2):647–662, 2004.

[193] Bill Lydon. Internet of things industrial automation industry exploring and implementing IoT. *InTech Magazine*, March-April 2014.

[194] Yuxin Ma, Hongbo Shi, Hehe Ma, and Mengling Wang. Dynamic process monitoring using adaptive local outlier factor. *Chemometr Intell Lab Syst*, 127(0):89 – 101, 2013.

[195] J.F. MacGregor and T. Kourti. Statistical process control of multivariate processes. *Control Eng Pract*, 3(3):403 – 414, 1995.

[196] J. Makhoul. Linear prediction: A tutorial review. In *Proceedings of the IEEE*, volume 63, pages 561–580, April 1975.

[197] Stéphane Mallat. *A wavelet tour of signal processing*. The Sparse Way: Academic Press, 3rd edition, 2008.

[198] C. L. Mallows. On some ttopic in robustness. Technical report, Bell Telephone Laboratories Technical Memorandum, Murray Hill, New Jersey,, 1975.

[199] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity*. The McKinsey Global Institute, McKinsey & Company, 2011. Accessed on October 7,2014.

[200] S. L. Marple. *Digitial spectral analysis with applications*. NJ: Prentice Hall, 1987.

[201] S. Lawrence Marple and Albert H. Nuttall. Experimental comparison of three multichannel linear prediction spectral estimators. *Communica-*

tions, Radar and Signal Processing, IEE Proceedings F, 130(3):218–229, 1983.

[202] S. Marsland. *on-line novelty detection through self-organization, with application to Inspection robotics*. PhD thesis, University of Manchester, 2001.

[203] Harold Martens and Tormod Næs. *Multivariate calibration*. New Jersey: John Wiley & Sons, Ltd., 1989.

[204] R. Douglas Martin and Victor J. Yohai. Influence functionals for time series. *Ann Stat*, 14(3):781–818, 1986.

[205] R.D. Martin and D.J. Thomson. Robust-resistant spectrum estimation. *Proc IEEE*, 70(9):1097–1115, Sept 1982.

[206] Karen F. McBrayer and Thomas F. Edgar. Bias detection and estimation in dynamic data reconciliation. *J Process Control*, 5(4):285 – 289, 1995.

[207] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometr Intell Lab Syst*, 118(0):62 – 69, 2012.

[208] Yu Miao, Hongye Su, Wei Wang, and Jian Chu. Simultaneous data reconciliation and joint bias and leak estimation based on support vector regression. *Comput Chem Eng*, 35(10):2141 – 2151, 2011.

[209] Aleksandar D. Micić and Miroslav R. Mataušek. Optimization of PID controller with higher-order noise filter. *J Process Control*, 24(5):694 – 700, 2014.

[210] Michael Misiti, Yves Misiti, Georges Oppenheim, and Jean-Michel Poggi. *Wavelet Toolbox*. The MathWorks, Inc, 1996.

[211] Thomas M. Mitchell. *Machine learning*. New York: McGraw-Hill,, 1st edition, 1997.

[212] Alberto Muñoz and Jorge Muruzábal. Self-organizing maps for outlier detection. *Neurocomputing*, 18(1C3):33 – 60, 1998.

[213] Jose Co Munoz and Junghui Chen. Removal of the effects of outliers in batch process data through maximum correntropy estimator. *Chemometr Intell Lab Syst*, 111(1):53 – 58, 2012.

[214] Koji Muteki, John F. MacGregor, and Toshihiro Ueda. Estimation of missing data using latent variable methods with auxiliary information. *Chemometr Intell Lab Syst*, 78(1-2):41 – 50, 2005.

[215] Alexandre Nairac, Neil Townsend, Roy Carr, Steve King, Peter Cowley, and Lionel Tarassenko. A system for the analysis of jet engine vibration data. *Integr Comput Aided Eng*, 6(1):53–66, 1999.

[216] Shankar Narasimhan and Cornelius Jordache. *Data reconciliation and gross error detection*. Gulf Professional Publishing, Burlington: TX, 1999.

[217] Mary Natrella. *e-Handbook of statistical methods*. NIST/SEMATECH, July 2010.

[218] Philip R.C. Nelson. *The treatment of missing measurements in PCA and PLS models*. PhD thesis, McMaster University,Hamilton, Ontario, Canada,, 2002.

[219] Philip R.C. Nelson, Paul A. Taylor, and John F. MacGregor. Missing data methods in pca and pls: Score calculations with incomplete observations. *Chemometr Intell Lab Syst*, 35(1):45 – 65, 1996.

[220] Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.*, 27(1):27–57, 2001.

[221] Boyi Ni, Deyun Xiao, and Sirish L. Shah. Time delay estimation for MIMO dynamical systems c with time-frequency domain analysis. *J Process Control*, 20(1):83 – 94, 2010.

[222] Niels-Peter Vest Nielsen, Jens Michael Carstensen, and Jø rn Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A*, 805(1-2):17–35, May 1998.

[223] Albert H Nuttall. Fortran program for multivariate linear predictive spectral analysis, employing forward and backward averaging. Technical report, DTIC Document, 1976.

[224] Albert H Nuttall. Multivariate linear predictive spectral analysis employing weighted forward and backward averaging: A generalization of burg's algorithm. Technical report, DTIC Document, 1976.

[225] Takayuki Okatani and Koichiro Deguchi. On the wiberg algorithm for matrix factorization in the presence of missing components. *Int J Comput Vis*, 72(3):329–337, 2007.

[226] Takayuki Okatani, Takahiro Yoshida, and Koichiro Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms. In *International Conference on Computer Vision*, pages 842–849, 2011.

[227] Georges Oppenheim, Anne Philippe, and Jean de Rigal. The particle filters and their applications. *Chemometr Intell Lab Syst*, 91(1):87 – 93, 2008.

[228] Sophocles J. Orfanidis. *Introduction to signal processing*. NJ: Prentice Hall,, 1996.

[229] Derya B. Özyurt and Ralph W. Pike. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Comput Chem Eng*, 28(3):381 – 402, 2004.

[230] R.K. Pearson. Exploring process data. *J Process Control*, 11(2):179 – 194, 2001.

[231] R.K. Pearson. Outliers in process modeling and identification. *IEEE Trans Control Syst Technol*, 10(1):55–63, Jan 2002.

[232] Randy J. Pell. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometr Intell Lab Syst*, 52(1):87 – 104, 2000.

[233] Daniel Peña. Influential observations in time series. *J Bus Econ Stat*, 8(2):235–241, 1990.

[234] Greet Pison and Stefan Van Aelst. Analyzing data with robust multivariate methods and diagnostic plots. In Wolfgang Härdle and Bernd Rónz, editors, *Compstat*, pages 165–170. Heidelberg: Physica-Verlag, 2002.

[235] Amogh V. Prabhu, Thomas F. Edgar, and Rick Good. Missing data estimation for run-to-run ewma-controlled processes. *Comput Chem Eng*, 33(11):1861 – 1869, 2009.

[236] J. Prakash, Biao Huang, and Sirish L. Shah. Recursive constrained state estimation using modified extended kalman filter. *Comput Chem Eng*, 65(0):9 – 17, 2014.

[237] Eranda Harinath Puwakkatiya-Kankanamage, Salvador García-Muñoz, and Lorenz T. Biegler. An optimization-based undeflated pls (oupls) method to handle missing data in the training set. *J. Chemom.*, 28(7):575–584, 2014.

[238] Joe S. Qin. Process data analytics in the era of big data. *AIChE J.*, 60:3092–3100, 2014.

[239] S. Joe Qin, Sergio Valle, and Michael J. Piovoso. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemom.*, 15(9):715–742, 2001.

[240] J. Ross Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Series in Machine Learning. San Mateo: Morgan kaufmann,, 1st edition, September 1993.

[241] J.R. Quinlan. Induction of decision trees. *Mach Learn*, 1(1):81–106, 1986.

[242] Lawrence R Rabiner and Bernard Gold. *Theory and application of digital signal processing*. Englewood Cliffs, N.J.:Prentice-Hall, Inc., 1975.

[243] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 427–438, New York, NY, USA, 2000. ACM.

[244] Mark R. Raymond and Dennis M. Roberts. A comparison of methods for treating incomplete data in selection research. *Educ Psychol Meas*, 47(1):13–26, 1987.

[245] X.M. Ren, A.B. Rad, P.T. Chan, and W.L. Lo. Online identification of continuous-time systems with unknown time delay. *IEEE Trans Automat Contr*, 50(9):1418–1422, Sept 2005.

[246] Douglas Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, pages 659–663. Springer,, 2009.

[247] Jean-Pierre Richard. Time-delay systems: an overview of some recent advances and open problems. *Automatica*, 39(10):1667 – 1694, 2003.

[248] S. J. Roberts. Novelty detection using extreme value statistics. *Vision, Image and Signal Processing, Proc IEE*, 146(3):124–129, Jun 1999.

[249] Stephen Roberts and Lionel Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Comput.*, 6(2):270–284, 1994.

[250] B. Roffel and B. Betlem. *Advanced practical process control*. Advances in soft computing. Springer,, 2004.

[251] Lior Rokach and Oded Maimon. *Data mining with decision trees: theory and applications*, volume 81 of *Series in Machine Perception and Artificial Intelligence*. Singapore: World Scientific, 2nd edition, 2014.

[252] Frank Rosenblatt. *Principles of neurodynamics, perceptrons and the theory of brain mechanisms*. Washington DC: Spartan Books,, 1st edition, 1961.

[253] Philip L. Roth. Missing data: a conceptual review for applied psychologists. *Pers Psychol*, 47(3):537–560, 1994.

[254] Peter Rousseeuw and Victor Yohai. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. New York:Springer-Verlag, 1984.

[255] Peter J. Rousseeuw. Least median of squares regression. *J. Am. Stat. Assoc.*, 79(388):871–880, 1984.

[256] Peter J. Rousseeuw. Multivariate estimation with high breakdown point. *Math Stat Appl*, B:283–297, 1985.

[257] Peter J. Rousseeuw and Annick M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.,, 3rd edition, 1996.

[258] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[259] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley Series in probability and Mathematical statistics. New Jersey: John Wiley & Sons, Ltd.,, 1st edition, 1987.

[260] Evan L. Russell, Leo H. Chiang, and Richard D. Braatz. Fault detection in industrial processes using canonical variate analysis and dynamic

principal component analysis. *Chemometr Intell Lab Syst*, 51(1):81 – 93, 2000.

[261] Stuart Russell and Peter Norvig. *Artificial intelligence: A modern approach*. NJ: Prentice Hall, 3rd edition, December 2009.

[262] Tito L.M. Santos, Paulo E.A. Botura, and Julio E. Normey-Rico. Dealing with noise in unstable dead-time process control. *J Process Control*, 20(7):840 – 847, 2010.

[263] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[264] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC Monographs on Statistics & Applied Probability. Florida: Chapman & Hall/ CRC,, 1st edition, August 1997.

[265] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Pyschol Methods*, 7(2):147–177, 2002.

[266] Alois Schlögl. A comparison of multivariate autoregressive estimators. *Signal Processing*, 86(9):2426–2429, 2006.

[267] Tapio Schneider and Arnold Neumaier. Algorithm 808: Arfit - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.*, 27(1):58–65, 2001.

[268] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[269] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[270] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[271] Dale E Seborg, Duncan A Mellichamp, Thomas F Edgar, and Francis J Doyle III. *Process dynamics and control*. John Wiley & Sons,, 3rd edition, 2010.

[272] V. Romero Segovia, T. Hägglund, and K.J. Åström. Measurement noise filtering for {PID} controllers. *J Process Control*, 24(4):299 – 313, 2014.

[273] Sven Serneels and Tim Verdonck. Principal component analysis for data containing outliers and missing elements. *Comput Stat Data Anal*, 52(3):1712 – 1727, 2008.

[274] Anil K. Seth. A MATLAB toolbox for granger causal connectivity analysis. *J. Neurosci. Meth.*, 186(2):262 – 273, 2010.

[275] Shashi Shekhar, Chang-Tien Lu, and Pusheng Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proceedings of the seventh ACM SIGKDD international con-*

ference on Knowledge discovery and data mining, pages 371–376. ACM, 2001.

[276] Hailin Shen, Glyn Nelson, Stephnie Kennedy, David Nelson, James Johnson, David Spiller, Michael R.H. White, and Douglas B. Kell. Automatic tracking of biological cells and compartments using particle filters and active contours. *Chemometr Intell Lab Syst*, 82(1C2):276 – 282, 2006.

[277] Erich Shubert, Remigius Wojdanowski, Hans-Peter Kriegel, and Arthur Zimek. On evaluation of outlier rankings and outlier scores. In *Proceedings of 12th SIAM International Conference on Data Mining*, 2012.

[278] Heung-Yeung Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Trans Pattern Anal Mach Intell*, 17(9):854–867, Sep 1995.

[279] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, Manuel López-Coello, and María-Dolores Cubiles de-la Vega. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Netw*, 24(1):121 – 129, 2011.

[280] Anita Singh. Outliers and robust procedures in some chemometric applications. *Chemometr Intell Lab Syst*, 33(2):75 – 100, 1996.

[281] Tyler Soderstrom. *Integration of on-line data reconciliation and bias identification techniques*. PhD thesis, The University of Texas at Austin, Austin, TX, USA,, 2001.

[282] Tyler A. Soderstrom, David M. Himmelblau, and Thomas F. Edgar. A mixed integer optimization approach for simultaneous data reconciliation and identification of measurement bias. *Control Eng Pract*, 9(8):869 – 876, 2001.

[283] David S. Stoffer. Detecting common signals in multiple time series using the spectral envelope. *J. Am. Statist. Assoc.*, 94:94–1341, 1998.

[284] David S. Stoffer, David E. Tyler, and Andrew J. McDougall. Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80(3):pp. 611–622, 1993.

[285] David S. Stoffer, David E. Tyler, and David A. Wendt. The spectral envelope and its applications. *Stat. Sci.*, 15(3):pp. 224–253, 2000.

[286] Maria Gabriella Tana, Roberta Sclocco, and Anna Maria Bianchi. Gmac: A matlab toolbox for spectral granger causality analysis of fmri data. *Comput. Biol.Med.*, 42(10):943 – 956, 2012.

[287] Jian Tang, Zhixiang Chen, Ada Wai chee Fu, and David Cheung. A robust outlier detection scheme for large data sets. In *In 6th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 6–8, 2001.

[288] Jian Tang, Zhixiang Chen, Ada Wai-Chee Fu, and David W Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Advances in Knowledge Discovery and Data Mining*, pages 535–548. Berlin Heidelberg:Springer, 2002.

[289] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach Learn*, 54(1):45–66, January 2004.

[290] Ming T. Tham, Gary A. Montague, A. Julian Morris, and Paul A. Lant. Soft-sensors for process estimation and inferential control. *J Process Control*, 1(1):3 – 14, 1991.

[291] Nina F. Thornhill, John W. Cox, and Michael A. Paulonis. Diagnosis of plant-wide oscillation through data-driven analysis and process understanding. *Control Eng. Pract.*, 11(12):1481 – 1490, 2003.

[292] Nina F. Thornhill and Alexander Horch. Advances and new directions in plant-wide disturbance detection and diagnosis. *Control Eng. Pract.*, 15(10):1196 – 1206, 2007. ADCHEM 2006 Special Issue.

[293] A. N. Tikhonov and V.Y. Arsenin. *Solution of ill-posed problems*. Winston & Sons, Washington, 1977.

[294] I.B. Tjoa and L.T. Biegler. Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. *Comput Chem Eng*, 15(10):679 – 690, 1991.

[295] A J Toprac, D J Downey, and S Gupta. Run-to-run control process for controlling critical dimensions, 1999.

[296] Philip H. S. Torr and David W. Murray. Outlier detection and motion segmentation. *Proc SPIE*, 2059:432–443, 1993.

[297] Ruey S. Tsay. Outliers, level shifts, and variance changes in time series. *J Forecasting*, 7(1):1–20, 1988.

[298] Ruey S. Tsay, Daniel Peña, and Alan E. Pankratz. Outliers in multivariate time series. *Biometrika*, 87(4):789–804, 2000.

[299] Nikos Tsikriktsis. A review of techniques for treating missing data in om survey research. *J Oper Manag*, 24(1):53 – 62, 2005.

[300] John W Tukey. *Exploratory data analysis*. Behavior Science. London:Pearson, 1st edition, 1977.

[301] John W. Turkey. *Exploratory data analysis*. Behavioral Science. London:Pearson, 1st edition, 1977.

[302] David A. van Dyk and Xiao-Li Meng. The art of data augmentation. *J Comput Graph Stat*, 10(1):1–50, 2001.

[303] P. Vankeerberghen, J. Smeyers-Verbeke, R. Leardi, C. L. Karr, and D. L. Massart. Robust regression and outlier detection for non-linear models using genetic algorithms. *Chemometr Intell Lab Syst*, 28(1):73 – 87, 1995.

[304] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, and H. Lä hdesmä ki. Self-organization and missing values in som and gtm. *Neurocomputing*, 147(0):60 – 70, 2015.

[305] Tommi Vatanen. Missing value imputation using subspace methods with applications on survey data. Master's thesis, Aalto University, Espoo, Finland,, 2012.

[306] Venkat Venkatasubramanian. Drowning in data: Informatics and modeling challenges in a data-rich networked world. *AIChE J.*, 55(1):2–8, 2009.

[307] Sabine Verboven and Mia Hubert. Libra: a MATLAB library for robust analysis. *Chemometr Intell Lab Syst*, 75(2):127 – 136, 2005.

[308] M. Vetterli and C. Herley. Wavelets and filter banks: theory and design. *IEEE Trans Signal Proc*, 40(9):2207–2232, Sep 1992.

[309] B. Walczak. Outlier detection in bilinear calibration. *Chemometr Intell Lab Syst*, 29(1):63 – 73, 1995.

[310] B. Walczak and D.L. Massart. Robust principal components regression as a detection tool for outliers. *Chemometr Intell Lab Syst*, 27(1):41 – 54, 1995.

[311] B. Walczak and D.L. Massart. Multiple outlier detection revisited. *Chemometr Intell Lab Syst*, 41(1):1 – 15, 1998.

[312] B. Walczak and D.L. Massart. Dealing with missing data: Part I. *Chemometr Intell Lab Syst*, 58(1):15 – 27, 2001.

[313] B. Walczak and D.L. Massart. Dealing with missing data: Part II. *Chemometr Intell Lab Syst*, 58(1):29 – 42, 2001.

[314] Richard Weber. Measurement smoothing with a nonlinear exponential filter. *AIChE J.*, 26(1):132–134, 1980.

[315] Peter D. Wentzell, Darren T. Andrews, David C. Hamilton, Klaas Faber, and Bruce R. Kowalski. Maximum likelihood principal component analysis. *J. Chemom.*, 11:339–366, 1997.

[316] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Analysis of multiblock and hierarchical pca and pls models. *J. Chemom.*, 12(5):301–321, 1998.

[317] D. Wettschereck. *A study of distance-based machine learning algorithms*. PhD thesis, Department of Computer Science, Oregon State University, Corvallis, 1994.

[318] T. Wiberg. Computation of principal components when data are missing. In *Symposium of Computational Statistics*, pages 229–236, 1976.

[319] Bernard Widrow and Eugene Walach. *Adaptive Inverse Control, Reissue Edition: A Signal Processing Approach*. John Wiley & Sons, Hoboken,New Jersey, 2008.

[320] Patrick Wiegand, Randy Pell, and Enric Comas. Simultaneous variable selection and outlier detection using a robust genetic algorithm. *Chemometr Intell Lab Syst*, 98(2):108 – 114, 2009.

[321] Nobert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT Press, 1st edition, 1964.

[322] G. Willems, G. Pison, P.J. Rousseeuw, and S. Van Aelst. A hotelling test based on MCD. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat*, pages 117–122. Heidelberg: Physica-Verlag, 2002.

[323] Barry M. Wise and Neal B. Gallagher. Multivariate modeling of batch processes using summary variables. Technical report, Eigenvector Research, Inc., Wenatchee, WA, October 2011.

[324] B.M. Wise, D.J. Veltkamp, N.L.Ricker, B.R. Kowalski, S.M. Barnes, and V. Arakali. Application of multivariate statistical process control (mspc) to the west valley slurry-fed ceramic melter process. In *Waste Management*, volume 2, Tucson, AZ, 1991.

[325] H. O. A. Wold. *A Study in the analysis of stationary time series*. A Study in the Analysis of Stationary Time Series. Almqvist & Wiksells boktrycheri-a.-b., 1938.

[326] Svante Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognit*, 8(3):127 – 139, 1976.

[327] Wongphaka Wongrat, Thongchai Srinophakun, and Penjit Srinophakun. Modified genetic algorithm for nonlinear data reconciliation. *Comput Chem Eng*, 29(5):1059 – 1067, 2005.

[328] Shu Xu, Michael Baldea, Thomas F. Edgar, Willy Wojsznis, Terrence Blevins, and Mark Nixon. An improved methodology for outlier detection in dynamic datasets. *AIChE J.*, 61(2):419–433, 2015.

[329] Xuefeng Yan. Multivariate outlier detection based on self-organizing map and adaptive nonlinear map and its application. *Chemometr Intell Lab Syst*, 107(2):251 – 257, 2011.

[330] Fan Yang, Ping Duan, Sirish L. Shah, and Tongwen Chen. *Capturing Connectivity and Causality in Complex Industrial Processes*. SpringerBriefs in Applied Sciences and Technology. Springer International Publishing, 1 edition, 2014.

[331] Fan Yang, Sirish Shah, and Deyun Xiao. Signed directed graph based modeling and its validation from process knowledge and process data. *Int. J. Appl. Math. Comput. Sci.*, 22(1):41–53, 2012.

[332] Zi-Jiang Yang, Tomohiro Hachino, and Teruo Tsuji. On-line identification of continuous time-delay systems combining least-squares techniques with a genetic algorithm. *Int J Control*, 66(1):23–42, 1997.

[333] Jie Yu and S. Joe Qin. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J.*, 54(7):1811–1829, 2008.

[334] Qi Yu, Yoan Miche, Emil Eirola, Mark van Heeswijk, Eric Sé verin, and Amaury Lendasse. Regularized extreme learning machine for regression

with missing data. *Neurocomputing*, 102(0):45 – 51, 2013. Advances in Extreme Learning Machines (ELM 2011).

[335] Tao Yuan and S. Joe Qin. Root cause diagnosis of plant-wide oscillations using granger causality. *J. Process Contr.*, 24(2):450 – 459, 2014. ADCHEM 2012 Special Issue.

[336] Jiusun Zeng and Chuanhou Gao. Improvement of identification of blast furnace ironmaking process by outlier detection and missing value imputation. *J Process Control*, 19(9):1519 – 1528, 2009.

[337] S. Zhang, Zhenxing Qin, C. X. Ling, and S. Sheng. "missing is useful": missing values in cost-sensitive decision trees. *IEEE Trans Knowl Data Eng*, 17(12):1689–1693, Dec 2005.

[338] Xiao-Yu Zhang, Qing-Bo Li, and Guang-Jun Zhang. Modified robust continuum regression by net analyte signal to improve prediction performance for data with outliers. *Chemometr Intell Lab Syst*, 107(2):333 – 342, 2011.

[339] Yushi Zhang and Waleed H. Abdulla. A comparative study of time-delay estimation techniques using microphone arrays. Technical Report 619, Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand,, 2005.

[340] Zhengdao Zhang and Feilong Dong. Fault detection and diagnosis for

missing data systems with a three time-slice dynamic bayesian network approach. *Chemometr Intell Lab Syst*, 138(0):30 – 40, 2014.

[341] Zhengjiang Zhang and Junghui Chen. Simultaneous data reconciliation and gross error detection for dynamic systems using particle filter and measurement test. *Comput Chem Eng*, 69(0):66 – 74, 2014.

[342] Zhonggai Zhao, Biao Huang, and Fei Liu. Bayesian method for state estimation of batch process with missing data. *Comput Chem Eng*, 53(0):14 – 24, 2013.

[343] Zhonggai Zhao, Biao Huang, and Fei Liu. Parameter estimation in batch process using EM algorithm with particle filter. *Comput Chem Eng*, 57(0):159 – 172, 2013. PSE-2012.

[344] Zhonggai Zhao, Qinghua Li, Min Huang, and Fei Liu. Concurrent pls-based process monitoring with incomplete input and quality measurements. *Comput Chem Eng*, 67(0):69 – 82, 2014.

[345] D Zhen, H L Zhao, F Gu, and A D Ball. Phase-compensation-based dynamic time warping for fault diagnosis using the motor current signal. *Meas. Sci. Technol.*, 23(5):55601, 2012.

[346] D. H. Zhou and P. M. Frank. A real-time estimation approach to time-varying time delay and parameters of NARX processes. *Comput Chem Eng*, 23(11C12):1763 – 1772, 2000.

[347] J. Zhou and R.H. Luecke. Estimation of the covariances of the process noise and measurement noise for a linear discrete dynamic system. *Comput Chem Eng*, 19(2):187 – 195, 1995.

[348] Xi Yu Zhou and Joon S. Lim. Replace missing values with EM algorithm based on GMM and Naive Bayesian. *Int J Soft Eng Res Appl*, 8(5):177–188, 2014.

[349] Jinlin Zhu, Zhiqiang Ge, and Zhihuan Song. Robust modeling of mixture probabilistic principal component analysis and process monitoring application. *AIChE J.*, 60(6):2143–2157, 2014.

[350] Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, and George Lapis. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. New York:McGraw-Hill Osborne Media,, 2011.

Vita

Shu Xu was born in Jintan, Jiangsu Province, China. He received his Bachelor of Engineering degree in Chemical Engineering from Tianjin University in July, 2011. In the Fall of 2011, he began doctoral study under the direction of Thomas F. Edgar at the University of Texas at Austin.

Permanent address: richard041123@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.